STATISTICAL MODELLING OF MOSQUITO ABUNDANCE AND WEST NILE VIRUS RISK WITH WEATHER CONDITIONS

YURONG CAO

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

April 2017

Abstract

Weather affects the abundance of mosquito vectors of mosquito-borne infectious diseases such as West Nile virus (WNv). Study and prediction of these effects could be used to develop disease forecasting methods. In this dissertation, we analyzed the frequency distribution of mosquito surveillance data and built the statistical forecasting models to predict the West Nile virus risk. In the first part, using mosquito data from the surveillance program in Peel Region, Ontario, we studied the distribution properties of *Culex* mosquito abundance data for the period from 2004 to 2012. We first employed statistical clustering method to identify two clusters of mosquito traps. The validation against landuse data supported the hypothesis that the clustering result successfully captured the influence of geographic variation in habitat effects on mosquito abundance. Accounting for the occurrence of these clusters, distribution analysis showed that *Culex* mosquito abundance in Peel Region followed a gamma distribution. Further analysis showed that summer mean temperature has a significant effect on mosquito distribution properties. We defined a normal weather

threshold under which the mosquito abundance followed a gamma distribution and abnormal weather conditions under which the mosquito abundance deviated from a gamma distribution. A predictive statistical model by clusters to forecast mosquito abundance in Peel Region using weather conditions was developed. In the second part, we developed forecasting models to predict the *Culex* mosquito abundance, the WNv risk and human incidence in Great Toronto Area (GTA) under weather changes by model selection. The predictions were in a good agreement with the observations for the period from 2002 to 2012. The model selection was demonstrated to be an effective way to compare different models. In the final part, finite mixture model and Markov regression models were combined to develop model-based clustering with generalized linear regression to cluster time series. Quasi-likelihood approach was adopted to deal with the Markov chain in the data generating process and Estimation-Expectation algorithm was used to estimate the parameters. The proposed algorithm was tested on simulated data and applied to mosquito surveillance data in Peel Region.

Acknowledgements

First and foremost I offer my sincerest gratitude to my supervisor, Professor Huaiping Zhu. It has been a great honor to be his Ph.D. student. He is not only an academic advisor but also a mentor in my life. I could never finish this thesis without his warm encouragement, thoughtful guidance, critical comments, and correction. My co-supervisor, Professor Xin Gao has always been there to listen and give advice. I am deeply grateful to her for the long discussions that helped me sort out the technical details of my work. I am also thankful to her for carefully reading and commenting on countless revisions of this manuscript. Besides my supervisors, I would like to thank the rest of my Ph. D committee: Dr. Nicholas H. Ogden and Professor Steven Wang for their encouragement, insightful comments, and hard questions. I am grateful to York University and Department of Mathematics and Statistics for the financial support and help they provided throughout my graduate studies. I also would like to thank the members of LAMPS who have contributed immensely to my personal and professional time at York University. The group has

been a source of friendships as well as good advice and collaboration.

I cannot finish without thanking my family. For my parents, they raised me with love and supported me in all my pursuits. And most of the thanks to my loving, supportive and patient husband, Stanley, who cheered me up and stood by me through the good times and the bad. I also want to express my gratitude and deepest appreciation to my lovely sweet daughter, Olivia, for all the happiness she brings to me and for being a good girl.

Table of Contents

List of Tables

List of Figures

1 Introduction

West Nile Virus (WNv) is a mosquito-borne flavivirus typically transmitted between birds and mosquitoes, and could infect humans and other mammals. Approximately 80% of people who are infected with WNv are asymptomatic. Up to 20% of the people who become infected will display flu-like condition which is called West Nile fever whose symptoms include fever, headache, body aches, nausea, vomiting and sometimes swollen lymph glands or a skin rash on the chest, stomach, and back. These mild symptoms typically last for a few days. Less than 1% of the infections can result in neurological disorders known as West Nile encephalitis and West Nile meningitis, and if left untreated can result in death (CDC 2015).

The virus was first isolated from the serum of a febrile woman from the West Nile district of Uganda in 1937 (Smithburn et al. 1940). The first recognized epidemic of WNv occurred in Israel in 1951 where 123 of 303 residences had been infected (Bernkopf et al. 1953). Before mid-1990, there were rare large outbreaks of WNv and the epidemic had been restricted to Africa and Mediterranean area. Beginning

around 1996, WNv experienced a drastic resurgence. A large outbreak of WNv occurred in the urban area of Romania followed by the ones in Tunisia in 1997, Russia in 1999, Israel in 2000 and France in 2003 (Hayes et al. 2005). WNv is now considered to be an endemic pathogen globally.

In the Western Hemisphere, WNv was first detected in New York in 1999 (Lanciotti et al. 1999). WNv has quickly spread out from New York to the north, south, and west. Within five years, WNv spread westward across all 48 contiguous states, as well part of southern Canada, Mexico, and Central and South America (Reisen and Brault 2007). Between 1999 and 2012, over 37,088 human infection cases were reported to the Centers for Disease Control and Prevention (CDC) with 1,549 fatalities (CDC 2013), in which one of the worst epidemics occurred in 2012 with 5,674 WNv disease cases and 286 death.

The activity of WNv in Canada was first reported in birds and mosquitoes in 2001. The following year, 2002, has witnessed the biggest emergence of WNv in Ontario with 394 residents having laboratory evidence of WNv infection and there was also 20 cases detected in Quebec. The virus quickly spread westward into the prairie provinces in 2003: 947 confirmed cases in Saskatchewan, 144 in Manitoba , and 275 in Alberta. The virus was also detected in birds (but not in people) in Nova Scotia and New Brunswick by 2003.

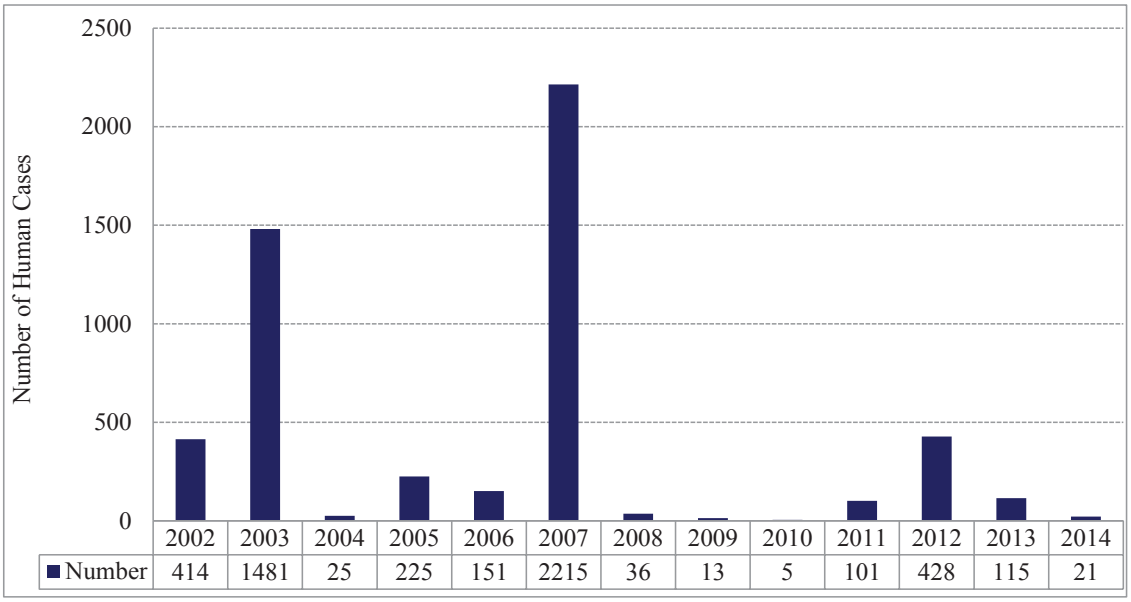

Figure 1.1 West Nile virus clinical cases and asymptomatic infections in Canada from 2002 to 2014. Data from the surveillance program of Public Health Agency of Canada.

Surveillance data from the Public Health Agency showed that there was no further geographical spread between 2004 and 2007 but the highest number of infections so far occurred in 2007 when a total of 2,215 human cases was reported and more than half were in Saskatchewan. The spread of WNv had eventually reached British Columbia while 2 locally acquired WNv was detected for the first time in 2009. Presently, it is reasonable to believe that the virus has established itself in North America and has returned to Canada, which was confirmed by the another outbreak in 2012 with 259 WNv human incidences being identified (PHAC 2015) (See Figure 1.1). There is no human vaccine currently available for WNv, which makes fore-

casting models and surveillance system important public health tools in the control and prevention of the disease. Therefore, understanding the population dynamics of mosquito and the WNv transmission under the impact of weather and environmental conditions is very important to inform implementation of control measures. In this dissertation, we studied the distribution properties of mosquito count in Peel Region, Ontario in order to investigate the association between mosquito abundance and weather factors within different types of landscape. Different statistical models were developed to predict the WNv vector mosquito abundance under weather conditions in Peel Region. Furthermore, we built statistical models to forecast WNv vector mosquito population, WNv risk and human incidence using Great Toronto Area (GTA) mosquito surveillance data under weather changes.

1.1 The WNv transmission cycle

The WNv survives by circulating between birds and mosquito populations. A female mosquito can acquire the infection by obtaining a blood meal from an infected bird and after a two-to-three week incubation period, can then pass the infection by injecting its saliva into another host (bird, horse, human or other animals) when it takes a blood meal. Once in the new host, the virus can multiply, causing illness and possibly death (PPH 2002). Mammals (such as horses) and humans are regarded as

dead-end hosts which are unable to uphold transmission cycles (Hayes et al. 2005) (see Figure 1.2).

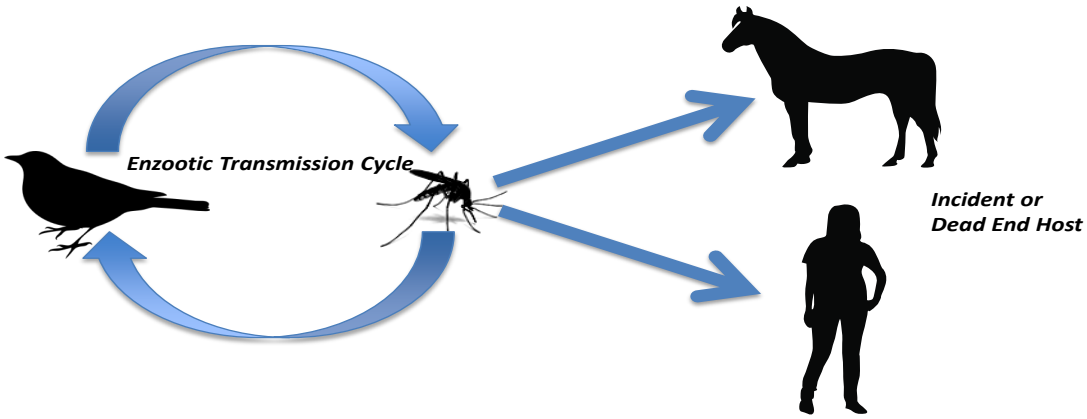

Figure 1.2 West Nile virus transmission cycle.

Culex pipiens and *Culex restuans* pose a very high risk for transmitting WNv to birds and humans in Ontario and have been estimated to be responsible for up to 80% of WNv human infection in the north eastern United states (Kilpatrick et al. 2005). The *Culex* species is often considered as an urban and suburban mosquito species because their common breeding sites are water habitats with a high organic content, often associated with storm water runoff commonly found in urban or suburban landscape (Kilpatrick et al. 2005). Since there is no vaccine available and clinical cure for WNv does not exit, reduction of mosquito populations remains the only way to reduce transmission of WNv. This makes forecasting models and surveillance systems an important public health tool in the control and prevention of the disease.

Therefore, understanding the population dynamics of mosquito and the WNv transmission under the impact of weather and environmental conditions is very important to inform implementation of control measures.

1.2 Statistical modeling of mosquito population and WNv risk

To develop effective control strategies, various statistical models have been built to forecast the mosquito population, WNv risk, and human incidence. Mosquito abundance and WNv transmission are both affected by weather and environmental factors, with temperature and precipitation considered key variables (Brown et al. 2008, Cooke et al. 2006, DeGroote et al. 2008, Pecoraro et al. 2007, Reisen et al. 2008). There have been different statistical modeling studies attempted to predict how climate change might affect the distribution of mosquito-borne diseases. These models fall into four primary categories: 1) multiple linear regression techniques, 2) time-series approaches, 3) generalized linear models, and 4)mixed effect models.

In 1986, Raddatz (1986) applied multiple linear regression techniques to seven years of data to generate a biometeorological model of Winnipeg's mean daily levels of *Culex arsalis Coquilletti*, a key vector for Western Equine Encephalitis. Pecoraro et al. (2007) discussed the climatic and landscape correlations for potential WNv

mosquito vectors in the Seattle region by multiple linear regression.

Time-series approaches are widely used to connect the mosquito population and virus transmission to weather conditions. An empirical model to forecast WNv mosquito vector populations in Erie County, New York was explored using time series analysis by Trawinski and Mackay (2008). Hu et al. (2004, 2006) developed epidemic, time-series forecasting models using local weather data and mosquito density to predict outbreaks of Ross River virus disease in Brisbane, Australia.

Since the mosquito vector has a skewed distribution (Costantino and Desharnais 1981, Hirano et al. 1982), generalized linear regression models are also applied by some researchers. Time-dependent Poisson regression model was developed to characterize the population dynamics of *Aedes sollicitans* (Walker) by using meteorological data and a 34-year set of daily mosquito count data in New Jersey by (Shone et al. 2006). Walsh et al. (2008) used Poisson regression to predict the seasonal abundance of mosquitoes in the same area based on off-season meteorological conditions. Costa et al. (2015) studied the relationship between egg number and climate and environmental variables through Bayesian zero-inflated spatial-temporal models. An and Rocklöv (2014) associated dengue fever in Hanoi with the meteorological determinants by negative binomial model.

Recently researchers used combination of the above methods due to the complex

biology of mosquitoes. Marcantonio et al. (2015) identified the environmental conditions favouring the WNv through linear mixed model. Yoo et al. (2016) used a linear mixed effects model with Poisson distribution to assess the effects of weather and landscape conditions on mosquito abundance. The association between mosquito abundance and weather and landscape conditions were analyzed by Rosa et al (Rosà et al. 2014)using generalized linear Poisson mixed effect model.

It is obvious that linear models are insufficient to replicate the complex biology of mosquitoes because of the skewed distribution of vector abundance. Time-series analysis itself also could not catch the variation of mosquito population under weather conditions. Some researchers used combination of the above methods and got better results, such as the work of Ruiz et al. (2010). They studied the local impact of temperature and precipitation on Minimum Infection Rate (MIR), an indicator of WNv infection, of *Culex* species mosquitoes in northeast Illinois using linear models with and without auto regression phase. The results showed that the model with auto regression is stronger. Another limitation was that those studies only employed one method to study the population of mosquito and transmission of the Vector-Borne Disease (VBD), and the method employed based on assumptions or previous research results in the other geographical regions. Five methods were tested by Abeku et al. (2002) to assess the accuracy of forecasting malaria incidence in area

with unstable transmission. Simple seasonal adjustment methods outperformed a statistically more advanced autoregressive integrated moving average method. The available studies and their results suggest that more accurate modeling using reliable predicting variables are essential, and when multiple models are used, model selection criteria should be employed to decide the best predictive model.

The recent work of Wang et al. (2011) discovered that weekly arithmetic means of mosquito counts of all traps in Peel region, Ontario follow a gamma distribution. A predictive statistical model for mosquito populations based on weather conditions was developed and optimism was provided for the development of weather-generated forecasting for WNv risk. Descloux et al. (2012) developed a climate-based multivariate non-linear statistical models using Support Vector Machines (SVM) technique to estimate the yearly risk of dengue outbreak in Noumena.

1.3 Overview of the dissertation

The overall goal of this thesis is to study the distribution properties of each individual trap in Peel Region in order to investigate the association between *Culex* mosquito abundance and weather factors within different types of landscape and to develop accurate temporal models to forecast WNv vector mosquito population, WNv risk and human incidence under weather changes.

We begin with Chapter 1 as the introduction and in Chapter 2, using mosquito data from the surveillance program of Ontario Ministry of Health and Long-Term Care (MOHLTC), we studied the distribution properties of *Culex pipens /restuans* mosquito abundance data in Peel Region, Ontario, Canada for the period from 2004 to 2012. We first employed statistical clustering to identify two clusters of mosquito traps and the validation against landuse data proved that the clustering result successfully captured the influence of geographic variation in habitat effects on mosquito abundance. Accounting for the occurrence of these clusters, distribution analysis showed that *Culex* mosquito abundance in Peel Region followed a gamma distribution. Further analysis showed that summer mean temperature has a significant effect on mosquito distribution properties. We defined a normal weather threshold under which the mosquito abundance followed a gamma distribution and abnormal weather conditions under which the mosquito abundance deviated from a gamma distribution. A predictive statistical model by clusters was developed to forecast the mosquito abundance using weather conditions.

In Chapter 3, we developed forecasting models to predict the *Culex* mosquito abundance, the WNv risk and human incidence in GTA under weather changes. We first examined the weather conditions that affect the mosquito abundance and WNv transmission, then gave the most significant temperature and precipitation conditions

in each case. Since the pattern of WNv and human cases are very complex, multiple models were employed to build the predictive models. Model selection criteria was used to choose the best fit models. The previous studies only chose the model by assumptions and did not verify it. By the model selection method used in this study, we have verified that the model we developed is the best fit model theoretically by a statistical method.

In Chapter 4, A mixture Markov regression model was proposed to analyze heterogeneous time series data. Mixture quasi-likelihood was formulated to model time series with mixture components and exogenous variables. The parameters were estimated by quasi-likelihood estimating equations. A modified EM algorithm was developed for the mixture time series model. The proposed algorithm was tested on simulated data set and applied to the analysis of the mosquito surveillance data in Peel Region.

2 Analysis of frequency distributions of mosquito surveillance data in Peel Region

2.1 Introduction

Mosquitoes are the vectors of a wide range of human infectious diseases, such as Malaria, Dengue, Yellow fever, Eastern Equine Encephalitis, St. Louis Encephalitis and WNv (CDC 2007). Mosquito abundance, mostly driven by climatic factors (temperature and precipitation affecting development, mortality and reproductive success), is a critical factor in outbreaks of mosquito-borne diseases (Patz et al. 1996). Consequently, understanding the effects of climatic factors on spatial and temporal dynamics of mosquito abundance is important for modelling and predicting the occurrence of vector-borne disease outbreaks and to inform implementation of control measures.

Statistical models have been widely used to predict the Vector-Bone disease and

virus transmission (Hu et al. 2006, Pecoraro et al. 2007, Raddatz 1986, Shone et al. 2006, Walsh et al. 2008).In many cases, however, precise prediction of mosquito abundance is impeded by the skewed distribution of mosquito abundance data. Ecological data such as nutrient concentrations, population densities, and biomasses are often lognormally distributed (Hirano et al. 1982, Singh et al. 1997). Costantino and Desharnais (1981) studied *Tribolium spp* beetle abundance and found the probability distribution of *Triboliumz* adults to be asymmetric and skewed to the right. They suggested that a gamma distribution was the best to describe their data. The recent work of Wang et al. (2011) discovered that weekly arithmetic means of mosquito counts of all traps in Peel region, Ontario, Canada also followed a gamma distribution. A predictive statistical model for mosquito abundance based on weather conditions was developed, and the optimism for the development of weather-generated forecasting of *Culex* mosquito abundance in Peel Region was provided. In their work (Wang et al. 2011), the arithmetic mean of the *Culex* mosquito counts per trap night was used, although environmental factors other than climate were not considered. The landscape can greatly affect mosquito abundance (Brownstein et al. 2002, Diuk-Wasser et al. 2006, Gómez et al. 2008, Pradier et al. 2008), therefore the predictions of the model built by Wang et al. (2011) may be coarse by not considering other environmental factors such as habitat.

In this chapter, we investigated the *Culex* mosquito abundance data from the surveillance program in Peel Region, Ontario and focused on the distribution properties of each individual trap in order to investigate the association between *Culex* mosquito abundance and weather factors including temperature and precipitation within different types of landscape. Peel region includes a diverse mixture of urban, suburban, rural, agricultural and natural landscapes; these geographic features vary across different trap locations. Using environmental data to characterise the trap sites first is a very good choice. However, environmental data are very rarely available that would allow such analysis and are rarely available at the fine geographic scale needed to precisely understand how they impact mosquito abundance responds to changes in weather variables. Therefore we first identified clusters of traps that yielded *Culex* mosquito abundance time series data that had similar patterns. Our hypothesis was that by so doing we would identify groupings of traps that had landscape/geographic features similar from the point of view of their effects on *Culex* mosquito reproduction and activity. We then applied the methods developed in Wang et al. (2011) to explore the distribution properties of *Culex* mosquito abundance data, and their variations due to weather, for each cluster of traps. A predictive statistical generalized linear model by clusters based on the distribution properties was developed for weather-based forecasting of the *Culex* mosquito abun-

dance in Peel Region. The model was calibrated and validated using actual weather and *Culex* mosquito data from the surveillance program in Peel Region, summer of 2012. The methods we proposed in this chapter will be used to further study the factors driving mosquito abundance in southern Ontario and other areas of Canada and improve the accuracy to predict how local weather and environmental factors influence *Culex* mosquito abundance and WNv risk.

2.2 Materials and Methods

2.2.1 Study site and program

Peel Region is situated in the west-central portion of the Greater Toronto Area (GTA), the largest urban agglomeration in Canada. It stretches from latitude 43^oN35 to 43^oN52. The most southern part of Peel is at longitude 79^oW37, while the most northern part of Peel is at longitude 80^oW0. Covering 1225 square kilometers (473 square miles), and stretching from Lake Ontario in the south to the Oak Ridges Moraine and the Niagara Escarpment to the north, Peel includes a diverse mixture of urban, suburban, rural, agricultural and natural landscapes. It comprises three municipalities: the cities of Mississauga, Brampton, and the town of Caledon. Peel Region has an estimated population of 1,296,814 based on 2011 Census data (Statistic Canada 2012).

WNv positive mosquitoes were first found in Peel in 2001 shortly after WNv-positive birds were discovered. The first full season of the mosquito surveillance program started in 2002. From 2003, working with the Ontario Ministry of Health and Long-Term Care (MOHLTC), the Ontario Ministry of Environment (MOE) and Health Canada, Region of Peel (the local public health unit) established a West Nile virus Prevention Plan. The plan comprises region-wide surveillance and mosquito control activities based on integrated pest management. The program emphasizes public education, source reduction, and larviciding. The mosquito data we have used in this study were collected via this program.

2.2.2 Data collection and processing

Since 2003, Centers for Disease Control (CDC) miniature light traps have been used to capture adult female mosquitoes in Peel Region. CDC light traps use carbon dioxide and light to attract female mosquitoes searching for a blood meal. Traps are set up once a week during the mosquito surveillance season, which lasts from June to September. Mosquitoes are collected from the traps the following morning, refrigerated and then transported on dry ice to a laboratory for identification. The mosquitoes are microscopically identified to species, counted and WNv vectors occurring in this region (*Cx. pipiens* and *Cx. restuans*) are tested for WNv (MOHLTC

2008, PPH 2008). In this study, we used mosquito surveillance data from 2004 to 2012 of 29 traps from which the data were collected each year and remained in the same location over this period. The data from 2004 to 2011 were used to perform clustering and distribution analysis, and the data of 2012 was used to validate the predictive model of *Culex* mosquito abundance. We used combined *Cx. pipiens* and *Cx. restuans* mosquito density data and the assumptions and caveats regarding the use of these combined abundance data (these species are difficult to separate in microscopic identification) are the same as those previously described (Wang et al. 2011). The numbers of mosquitoes captured per trap night, smoothed over proceeding and succeeding weeks, ($W_j = (w_{j-1}+w_j+w_{j+1})/3$), were used, where w_j was the original mosquito count in week j, and W_j was its smoothed value for that week. Smoothing was used to adjust for effects of moonlight on capture probabilities (Service 1993). The weather data used to analyze the climate impact on mosquito abundance were obtained from Canadas National Climate Archive (www.climate.weatheroffice.gc.ca). Data from the Pearson Airport weather station were used for analysis because this station has the longest and most complete record among the stations within and around Peel.

2.2.3 Cluster analysis

Generally, there are four major categories of clustering method: hierarchical methods, partitioning (nonhierarchical) algorithms, overlapping clustering procedures, and ordination techniques (Milligan and Cooper 1987). No definite rule indicates which type of clustering to use. K-means clustering belongs to the nonhierarchical method and is the most simple and famous algorithm. K-means and its variants have a time complexity that is linear in the number of documents, but are thought to produce inferior clusters (Steinbach et al. 2000). The most popular clustering algorithms have been the sequential agglomerative hierarchical methods (Lance and Williams 1967), although it has a limitation because of its quadratic time complexity. Let vector X_k represent the time series formed by the mosquito abundance in trap k from the year 2004 to 2011, where k varies from 1 to 29. We employed a hybrid clustering method suggested by Sclove (2001) to group this collection of time series in order to capture the influence of geographic variation in habitat effect on mosquito abundance. Non-hierarchical K-means method was combined with an agglomerative hierarchical method to obtain the clustering result (Steinbach et al. 2000). The clustering procedure started with a hierarchical algorithm to generate the initial clusters. The centroid of each cluster that was produced from the cluster dendrogram was utilized as the starting position of the centroid for input in the K-means method. Then

the K-means method was applied to obtain the cluster memberships. The number of clusters was determined by minimizing the Bayesian information criterion (BIC) (Schwarz et al. 1978):

$$BIC = n \times \log(WSS) + m \times \log(n), \tag{2.1}$$

where n is the number of observations. WSS is the total within-cluster sum of squares from points to the assigned cluster centers, and $m = k * p$ is the number of parameters, where k is the number of clusters, p is the length of a cluster center.

To validate whether or not identified clusters represented different environmental characteristics that may affect mosquito abundance, differences in landuse/landscape features between the clusters were investigated as follows. A circular buffer of two kilometers in radius (the approximate maximal dispersion distance for *Cx. pipens* (Lindquist et al. 1967, Moore et al. 1993)) was created around each trap location using ArcGIS 10.0 (ESRI). The area for each landuse/landscape type was calculated within each buffer zone. Landuse/landscape type data were obtained as vector files from the Credit Valley Conservation Foundation (http://www.creditvalleyca.ca/) by digitization of orthophotos acquired in 2011. Landuse variables were of two types, broad (urban, rural, natural) and detailed (53 classifications covering residential, construction, commercial/industrial, woodland types, open grassland types, agriculture types etc (Anderson 1976, NLCD 1992)). Associations between clusters and

broad and detailed landuse classification were investigated in simple logistic regression in STATA SE (version 11, STATACorp) with the cluster as the outcome variable and the area within the buffer for each landscape classification as the explanatory variable. To limit the possibility of type II errors, statistical analysis was limited to landscape classifications that comprised $> 2\%$ of the total landcover of the clusters. The level of significance was $P < 0.05$.

In order to incorporate the environmental factors into the clustering analysis, the environmental factors were combined with the mosquito abundance data to carry out the clustering analysis too. The procedure is as follows:

The same circular buffer was applied as defined above. The landuse/landscape type was grouped into three categories: build environment, Greenland and open area within each buffer zone. The indicator variables for the first two categories (which account for the majority in the buffer zone) and the elevation of each trap location were considered as environmental factors and were added as attributes into the clustering data set. Each attribute was scaled before clustering so all had variance equal to one. Since there were 128 mosquito abundance data (data is available once a week for 16 weeks each year over 8 years) and three environmental attributes (The environmental attributes do not change during the studying period), a bigger weight was needed to assign to the environmental factors. With the mosquito abundance data

assigned to a weight of 1, simulations were run with changing weight for environment attributes.

2.2.4 Distribution Analysis of *Culex* Mosquito Abundance

To explore the possible distributions of mosquito counts, we fitted the data using both gamma and lognormal distributions for each deduced cluster separately and performed hypothesis tests for the underlying distributions.

We considered a gamma distribution model which has the density function:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \tag{2.2}$$

where x is the mosquito count per trap night, α and β are the shape and scale parameters respectively. Wang et al. (2011) showed that the scale parameter β (i.e. the height of the curve) is mainly determined by temperature and precipitation, while the shape parameter α (which determines the degree of skewness) was assumed to be constant in every year over the study period. However, environmental factors can have an effect on both the shape and scale parameters (Atashi et al. 2009, Barker et al. 2009). In our analysis, we did not restrict the shape parameter and assumed the parameters α and β are functions of climate and environmental factors, which vary from year to year and cluster to cluster. We also considered a lognormal distribution,

which has the density function:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \tag{2.3}$$

where x is the mosquito count per trap night, the parameters μ and σ, are the mean and standard deviation respectively.

Using the VCD and STATS4 packages in the statistical software R, the parameters of the gamma and lognormal models for each cluster in different years were estimated. To assess the fit of the lognormal distribution, we applied log transformation of the data and performed Lilliefors test (Lilliefors 1967). Lilliefors test is an adaptation of the Kolmogorov-Smirnov test, but it is adjusted for the fact that the parameters could be estimated from the data rather than specified in advance, and it is particularly useful in the case of small samples. Although mosquito data in our study was not small, Lilliefors test is more appropriate while it is used in the other health regions where the surveillance data is not available in large amount. To assess the fit of a gamma distribution, the method of Wilding and Mudholkar (2008) was used. The gamma goodness-of-fit test statistic is:

$$z = \frac{1}{2} log(\frac{1-r}{1+r}), \tag{2.4}$$

where

$$r = \frac{\sum_{i=1}^{n}(\bar{x}_{-i} - \bar{x})(c_{-i} - \tilde{c})}{\sqrt{\sum_{i=1}^{n}(\bar{x}_{-i} - \bar{x})^2 (c_{-i} - \tilde{c})^2}}. \tag{2.5}$$

$x_1, x_2, \cdots, x_n, n \geq 3$ is a random sample, $\bar{x}$ and s are the sample mean and sample standard deviation, $c = s/\bar{x}$ is the sample coefficient of variation. n pairs $(\bar{x}_{-i}, c_{-i})$ are created by removing one observation at a time from the sample. $\tilde{c}$ is the mean of c_{-i}. For a gamma population, the test statistic z has an asymptotic normal distribution:

$$\sqrt{n}z \xrightarrow{distribution} N(0, 3 + \frac{10}{\alpha}), \tag{2.6}$$

where the shape parameter α is obtained from the previous gamma distribution fit. The level of significance was $P < 0.05$.

2.2.5 Modelling the impact of weather on *Culex* mosquito abundance

Culex mosquito abundance in Peel Region can be modeled by gamma distribution (Wang et al. 2011). We chose to use a generalized linear model (McCullagh and Nelder 1989) to build the predictive model by clusters for *Culex* mosquito abundance and the distribution property we obtained from the above analysis. Let $Y = (Y_i), i = 1, \cdots, n$, denote the *Culex* mosquito counts and X denotes the predictive variables, the generalized linear model is described by equation:

$$g(\mu) = X\alpha, \tag{2.7}$$

where g is called link function, $\mu = E(Y)$ is the mean of *Culex* mosquito counts, and α is the regression parameter vector. The canonical link function for gamma

distribution is reciprocal function $\eta(\mu) = \mu^{-1}$. Alternatively, log function $\eta(\mu) = \log \mu$ is also widely used. In this paper, both link functions were investigated.

We adopted the same method as in Wang et al. (2011) of growing degree days above 9^oC (dd) to show the impact of the temperature on mosquito abundance (Madder et al. 1983). It is defined as follows:

$$dd = \begin{cases} 0^oC & T_m \leq 9^oC \\ T_m - 9^oC & T_m > 9^oC, \end{cases} \tag{2.8}$$

where T_m denotes the daily mean temperature. We calculated $ddm_k = \sum_{h=1}^{k} dd(h)/k$ as the arithmetic mean of daily dd over k days prior to collection of the surveillance datum. ddm_k values obtained when k varied from 1 to 60 days were explored as potential explanatory variables. We also calculated $ppm_k = \sum_{h=1}^{k} pp(h)/k$ as the arithmetic mean of daily precipitation pp over k days prior to collection of the surveillance data. ppm_k values when k varied from 1 to 60 days were explored as potential explanatory variables to predict *Culex* mosquito abundance. *Culex* mosquitoes typically overwinter as adult females in reproductive arrest (Nelms et al. 2013). Therefore the previous year's *Culex* mosquito average abundance was included as a potential predictive variable. The time-series of *Culex* mosquito counts indicated that the week number in the mosquito season had an effect on the mosquito population. The time-series analysis also showed the *Culex* mosquito abundance in Peel Region is a first order autoregressive process (Chuang et al. 2011, Simoes et al. 2013). Accord-

ingly we included the week number in the current mosquito season and the first order autoregressive term as potential explanatory variables. Combining all the predictive variables, the predictive models could be written as:

$$\begin{aligned}\log(\mu) = \quad & \alpha_0 + \alpha_1 ddm_k + \alpha_2 ppm_l + \alpha_3 ddm_k \times ppm_l + \alpha_4 ddm_k^2 + \alpha_5 ppm_l^2 + \\ & \alpha_6 t + \alpha_7 t^2 + \alpha_8 M_p + \alpha_9 AR,\end{aligned} \tag{2.9}$$

where μ is the weekly mean mosquito abundance of the cluster, t is the current number of week in the mosquito season, M_p is the mean mosquito counts of last year, AR is the first order auto regression term of mosquito abundance and αs are the regression coefficients.

Several models were considered in this study. The first model included only temperature and precipitation as predictive variables. The second model added week number in the mosquito season as the independent variable. The third model was expanded to include the mean mosquito abundance of the previous year as a predictive variable. The fourth model included the first order autoregressive term. In each case, both of the two link functions, reciprocal, and log functions, were compared. In total, 8 models were investigated. Leave-one-out cross-validation was used as model validation. Among the data from 2004 to 2011, the data was partitioned with 7-year data as training data and the remaining one year as testing data. The cross-validation process was then repeated 8 times so that the data of each year was used exactly once as the validation data. The best fit model was determined by the

coefficient of determination, R^2, which is defined as:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}, \tag{2.10}$$

where n is the number of mosquito abundance in the validation data set, $SS_{err} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squares of residuals, $SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares, $\hat{y}_i$ is the predicted value of the ith observation, y_i is the corresponding observation, and $\bar{y}$ is the mean of all observed data. The model with the largest R^2 was chosen to be the best model for each cluster. The predictive model developed can be used to forecast the *Culex* mosquito abundance, and has been used to carry out the weekly forecasting of *Culex* mosquito abundance in the Region of Peel in 2012.

2.2.6 Impact of Weather on the Distribution Property

It was observed that most of the derivation between the theoretical and empirical distribution occurred at the tail part. To further investigate the climate factors which affect the distribution properties of the *Culex* mosquito abundance, we defined a deviation variable D as the difference between theoretical and the empirical tail probability of the data which falls beyond one standard derivation above the mean:

$$D = P_t(X \geq \mu + \sigma) - P_e(X \geq \mu + \sigma), \tag{2.11}$$

where P_t is the theoretical probability and P_e is the empirical probability. We raised the hypothesis that certain weather may affect mosquito abundance such that the capture data no longer follows a confirmed distribution. Usually, temperature and precipitation are the two important factors to determine the mosquito abundance.In the work of Trawinski and Mackay (2008), the weekly mean temperature and precipitation were used as two common variables to represent the impact of weather changes on mosquito abundance. In this paper, we used daily average temperature (in ^oC) and precipitation (in millimeters) (from May to September) in each year from 2004 to 2011 as explanatory variables. A multivariable linear-regression model was developed to explore the relationship between the dependent deviation variable D and the independent variables of temperature and precipitation.

2.3 Results

2.3.1 Clustering

The BIC was minimized at $k = 2$ (see Figure 2.1), so the traps were divided into two clusters. The time series plots for each cluster showed that the mosquito abundance of the traps in the same cluster covaried temporally suggesting a similar response to changes in climatic variables during the study period (see Figure 2.2).

Landscape variables were available for all 29 trap points. None of the broad

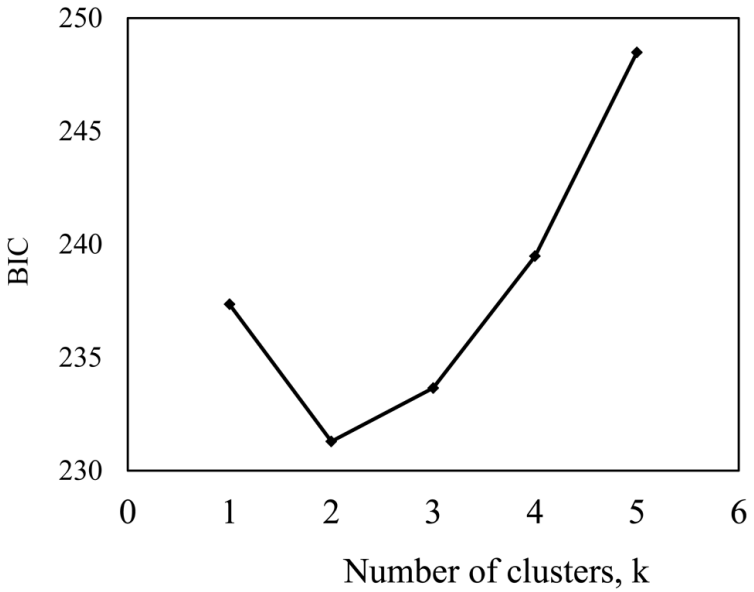

Figure 2.1 The BIC values for different numbers of clusters of mosquito data.

landscape classifications explained the separation of the trap points into clusters, however, trap points were significantly less likely to be in cluster 2 the greater was the area of uncultivated meadow around the trap, and marginally more likely the greater the area of deciduous woodland around the trap point (see Table 2.1 and Figure 2.3).

When the environmental factors weight ranged from 1 to 22, the clustering results including environmental factors remained the same. When the weight exceeded 23, the clustering result started to change. While the weight was equal to 22, the environmental factors contributed about 34% in the calculation of Euclidean distance in clustering. Having environmental attributes accounting for 34% of the information in

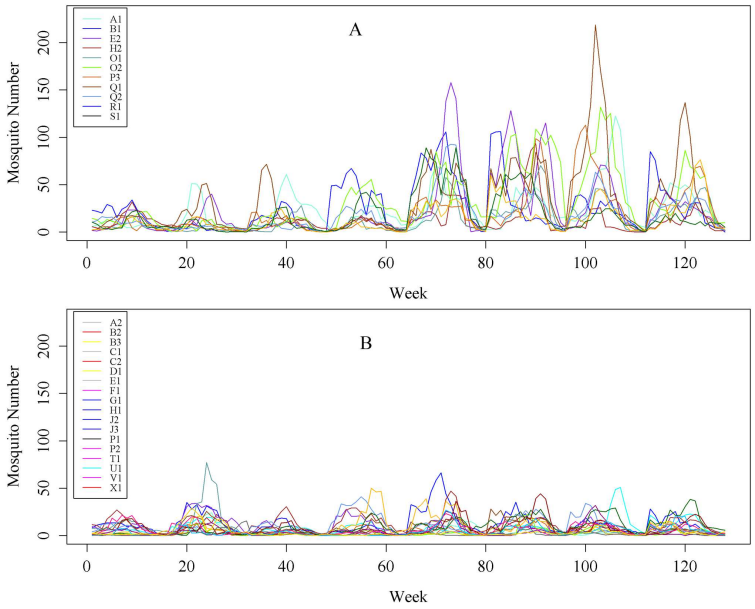

Figure 2.2 A time series plot of *Culex* mosquito abundance data in the two cluster of traps from 2004-2011. The symbols in the legend indicate the trap identification code. A) Cluster 1; B) Cluster 2.

clustering, the clustering result remained unchanged compared to the clustering without additional environmental attributes. This was a clear indication that clustering result without additional environmental attributes was robust to the accommodation of environmental factors. Furthermore, the mosquito abundance is affected by climate and environmental factors, with the temperature and precipitation being the key factors. It will downplay the effect of the weather factors if bigger weight is assigned to the environmental factors.

Landscape classification	% total landcover	Odds Ratio (95% CI)	Wald Z	P
Broad landscape classifications				
Natural	19.54	1 (0.99-1.00)	0.92	> 0.1
Urban	10.37	1 (0.99-1.00)	0.43	> 0.1
Rural	67.64	0.99 (0.99-1.00)	-0.99	> 0.1
Detailed landcover classifications				
Intensive agriculture	8.01	1 (0.99-1.00)	0.40	> 0.1
Commercial/industrial	11.08	0.99 (0.99-1.00)	-1.12	> 0.1
Uncultivated meadow	**6.43**	**0.99 (0.99-0.99)**	**-2.39**	**0.014**
Deciduous woodland	**4.05**	**1 (0.99-1.00)**	**1.84**	**0.065**
Recreational open space	4.27	0.99 (0.99-1.00)	-0.12	> 0.1
Open water	5.15	1 (0.99-1.00)	0.31	> 0.1
General urban area	11.31	0.99 (0.99-1.00)	-0.61	> 0.1
High density urban housing	2.69	1 (0.99-1.00)	0.79	> 0.1
Low density urban housing	22.03	1 (0.99-1.00)	0.40	> 0.1
Medium density urban housing	3.14	1 (0.99-1.00)	0.17	> 0.1

Table 2.1 The logistic regression analysis of associations between classification of trap sites into clusters and landscape variables.

2.3.2 Properties of *Culex* mosquito abundance

The density plots for each of the two clusters in different years from 2004 to 2011 showed that the mosquito counts had a distribution that was skewed to the left (see Figure 2.4). The results of the parameter estimation were listed in Table 2.2. The mean of shape parameters for cluster 1 and cluster 2 were 1.167 and 0.015 respectively (with standard derivations 0.331 and 0.147 respectively). The shape parameter β for cluster 2 varied little from year to year, but the shape parameter β for cluster 1 varied markedly.

The results of the goodness-of-fit test were shown in Table 2.3. If the P value is

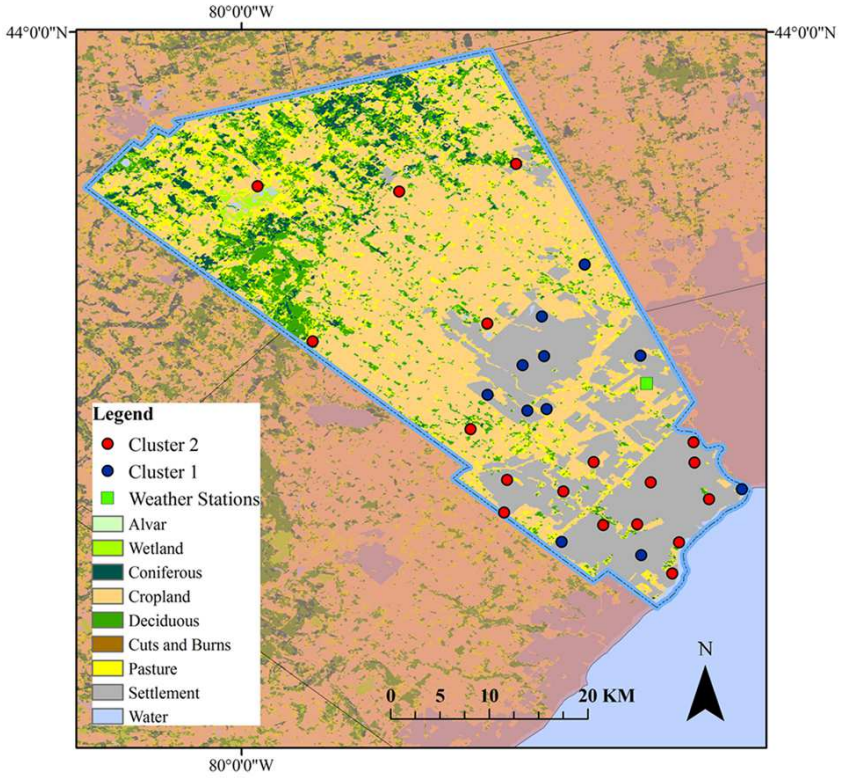

Figure 2.3 The location of the traps in the two different clusters on land use classification map (blue - cluster 1, red - cluster 2).

less than 0.05, we reject the null hypothesis that the data follows the stated distribution and accept the alternative hypothesis that the model and data are different. Most of the P values from the goodness-of-fit test of lognormal distribution were smaller than 0.05, with only 2 values above 0.05, so we concluded that the *Culex* mosquito abundance data did not not follow the lognormal distribution. Most of the P values from the goodness-of-fit test of the gamma distribution were much higher than 0.05, with the exception of that for cluster 1 in 2009 was higher than 0.05.

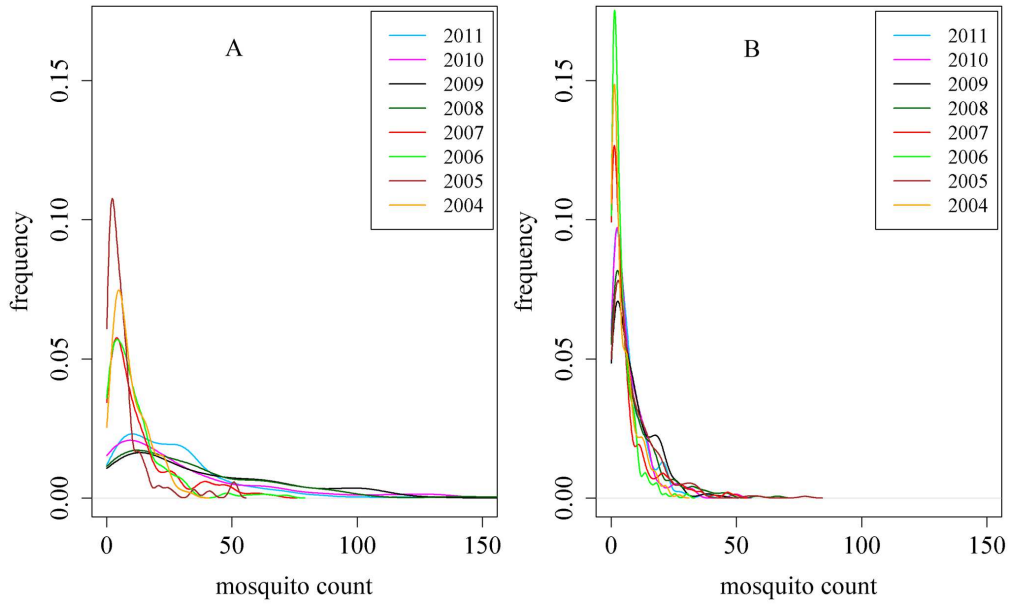

Figure 2.4 The frequency estimation of cluster 1 and cluster 2 from 2004-2011. A) Cluster 1; B) Cluster 2.

Based on these results, we concluded that the *Culex* mosquito abundance data followed the gamma distribution for cluster 2 and for most of the time for cluster 1, but it deviated from a gamma distribution under certain conditions.

2.3.3 Modelling the impact of temperature and precipitation

a) Most significant temperature and precipitation conditions

For cluster 1, ddm_{12} had the smallest P value of 0.000404 and the highest correlation coefficient of 0.3081 (Figure 2.5); for cluster 2, ddm_{12} also had the smallest P value of 4.297×10^{-13} and the highest correlation coefficient of 0.5847 (Figure 2.5). Therefore we chose ddm_{12} as the predictive variable for temperature for both cluster

	Gamma Parameter				Lognormal Parameter			
	Cluster 1		Cluster 2		Cluster 1		Cluster 2	
	Shape (α)	Scale (β)	Shape (α)	Scale (β)	meanlog	sdlog	meanlog	sdlog
2011	1.453	0.055	1.17	0.169	2.893	0.97	1.447	1.054
2010	0.837	0.027	1.085	0.15	2.737	1.373	1.451	1.099
2009	1.044	0.028	1.109	0.128	3.053	1.276	1.647	1.139
2008	1.099	0.033	0.908	0.101	2.991	1.209	1.55	1.233
2007	1.012	0.076	0.725	0.11	2.013	1.149	1.058	1.348
2006	1.044	0.092	1.126	0.273	1.879	1.182	0.912	1.048
2005	0.984	0.133	1.043	0.116	1.411	1.097	1.646	1.124
2004	1.863	0.196	0.952	0.196	1.961	0.809	0.969	1.171

Table 2.2 The estimated parameters of the gamma and lognormal distribution.

	P-value of Gamma		P-value of Lognormal	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
2011	0.7282	0.3387	0.0009	0.0022
2010	0.8973	0.7069	0.0035	0.0313
2009	0.0054	0.0908	0.0002	1.05E-05
2008	0.3603	0.8237	1.44E-05	0.0152
2007	0.2244	0.2398	0.6967	3.85E-06
2006	0.4436	0.8592	0.0016	0.0341
2005	0.3643	0.3692	0.2883	0.0061
2004	0.1729	0.0597	0.0204	3.21E-07

Table 2.3 The results of the Goodness of fit test.

1 and cluster 2. However, ppm_{35} had the smallest P value of and highest correlation coefficient of 0.5600 for cluster 1 (Figure 2.6), while for cluster 2, ppm_{30} had the smallest P value of 6.319×10^{-12} and highest correlation coefficient of 0.3922 (Figure 2.6). Hence, ppm_{35} was chosen for cluster 1 and ppm_{30} was chosen for cluster 2 as predictive variables for precipitation.

b) Best fit model for *Culex* mosquito abundance

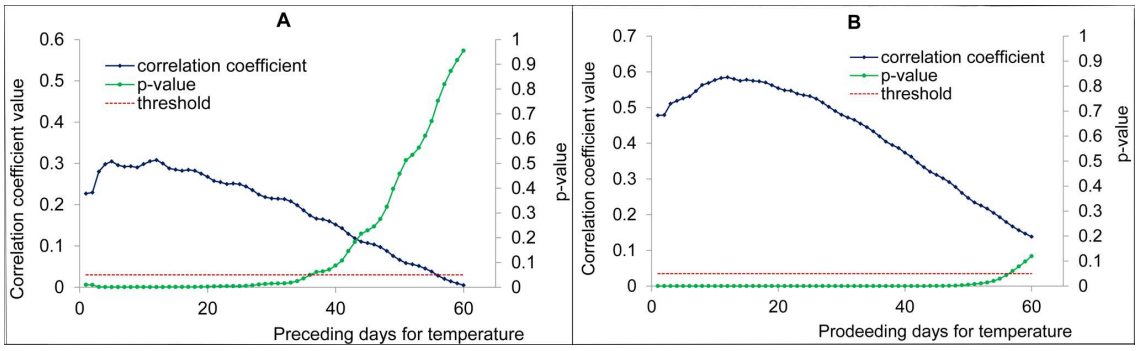

Figure 2.5 The correlation coefficients and P values between mosquito counts and different *ddm*, A) Cluster 1; B) Cluster 2.

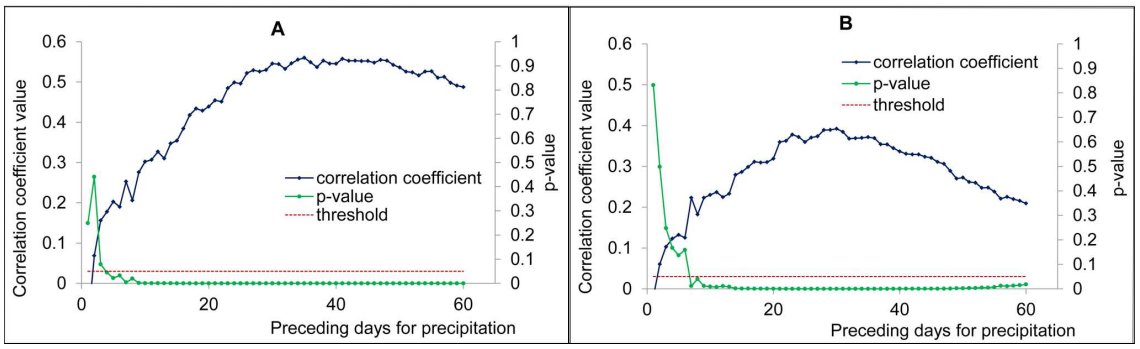

Figure 2.6 The correlation coefficients and P values between mosquito counts and different *ppm*, A) Cluster 1; B) Cluster 2.

After selecting the variables for temperature and precipitation, different models were explored to find the best-fit model for each cluster. The R^2 of each model for the two clusters were compared in Table 2.4. Among these models, model 8 achieved the largest R^2 of 0.5697 for cluster 1 and 0.9923 for cluster 2. Therefore model 8 was chosen to be the best-fit model to predict *Culex* mosquito in Peel Region for the two clusters and the corresponding regression coefficients were listed in Table 2.5.

c) Model validation

		Explanary Variable						
Model	Link Function	ddm	ppm	Time	Auto regressive term	Mean *Culex* abundance of previous year	R squared for cluster 1	R squared for cluster 2
1	reciprocal	Yes	Yes				0.2465	0.3906
2	reciprocal	Yes	Yes	Yes			0.2906	0.5391
3	reciprocal	Yes	Yes	Yes	Yes		0.3453	0.6102
4	reciprocal	Yes	Yes	Yes	Yes	Yes	0.3773	0.7514
5	Log	Yes	Yes				0.3275	0.8386
6	Log	Yes	Yes	Yes			0.3824	0.9842
7	Log	Yes	Yes	Yes	Yes		0.4304	0.9872
8	Log	Yes	Yes	Yes	Yes	Yes	0.5697	0.9923

Table 2.4 The R^2 of the predicting models for the two clusters in Peel Region.

	α0	α1	α2	α3	α4	α5	α6	α7	α8	α9
Cluster 1	2.0641	-0.0077	-0.0648	0.0018	-0.0004	0.0077	0.1269	-0.0129	0.0133	0.0329
Cluster 2	1.1607	-0.0206	0.1067	0.0005	0.0015	-0.0102	0.0539	-0.0081	-0.0189	0.1129

Table 2.5 The regression parameters of the mosquito predicting model of the two clusters in Peel Region.

Observed values for *Culex* mosquito abundance (both actual numbers of mosquitoes and seasonality) were well predicted by the models for both clusters (Figure 2.7 and Figure 2.8). For cluster 1, the model fit very well with the observation values for most of the years except 2009 and 2010. The time series of mosquito count in 2009 has two peaks and the model could predict the two peaks but overestimated the first peak. For the year 2010, the prediction agreed very well with the observation at the beginning and end of the mosquito season, but the predicted peak value was larger than the observed value. For cluster 2, the predicted values matched the observed values very well and the R^2 of the cluster 2 model reached 0.9923. Slight

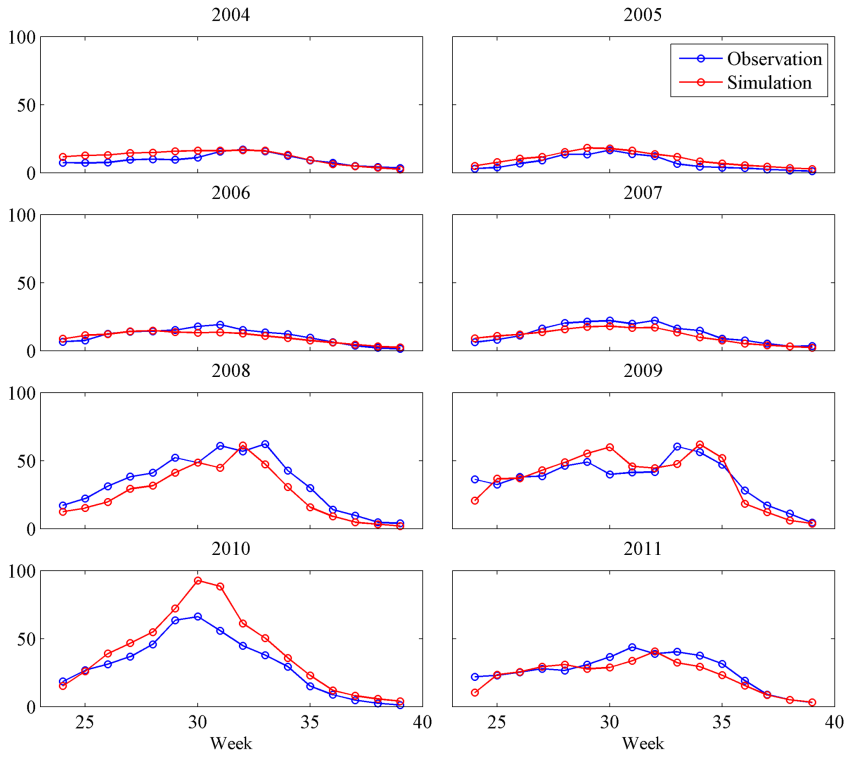

Figure 2.7 The observed versus model-predicted mosquito counts for traps in cluster 1.

over-prediction of peak mosquito counts were seen in years 2005, 2008 and 2010.

d) Model prediction simulation

The abundance of *Culex* mosquito predictive models developed with the data from 2004 through 2011 were validated against 2012 *Culex* mosquito season data by forecasting the 2012 data and then comparing with actual observed data. The mosquito forecasting program in Peel Region started in 2011 and continued in 2012. Every week in mosquito season (from the middle of June to earlier October), the mosquito traps were set up on Monday and Tuesday by the mosquito surveillance

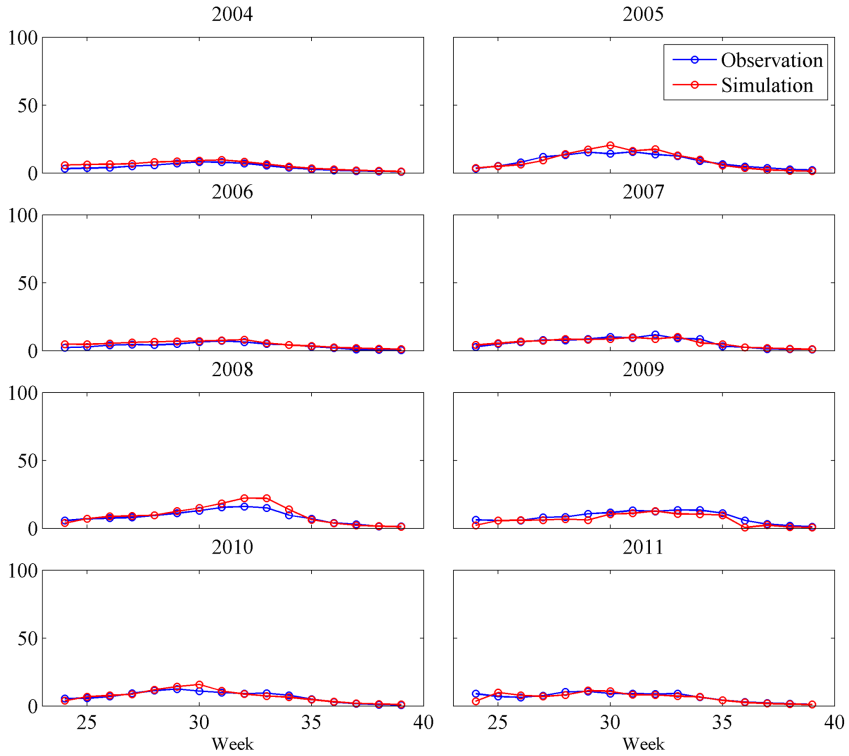

Figure 2.8 The observed versus model-predicted mosquito counts for traps in cluster 2.

program in Peel Region. The traps were collected the following morning and the mosquito data would be available on Wednesday. The previous weather data were collected through Canada's National Climate Archive (`http://climate.weatheroffice.gc.ca/Welcome_e.html`) and the weather data for the following two weeks were obtained through the weather forecasting network (`http://www.timeanddate.com/weather/canada/toronto/ext`). The mosquito predictive models would provide the *Culex* mosquito abundance data for the next two weeks by using the mosquito surveillance and weather data collected. The forecasting results were posted and updated

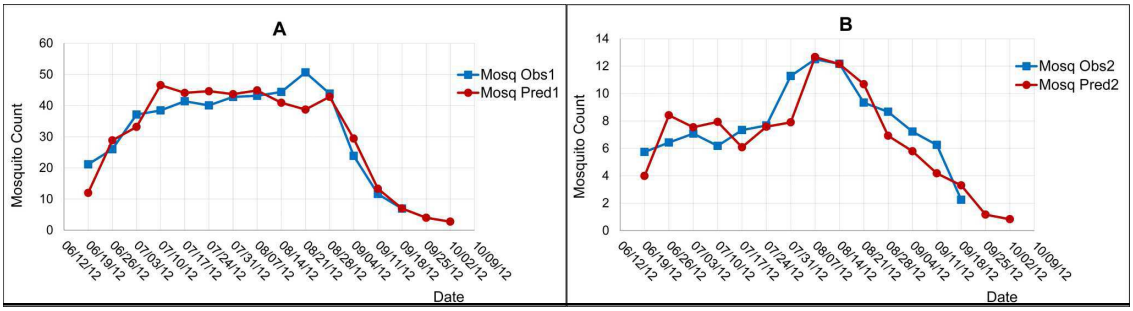

Figure 2.9 The observed versus the model-derived forecast of average mosquito counts per trap in the two clusters in 2012, A) Cluster 1; B) Cluster 2.

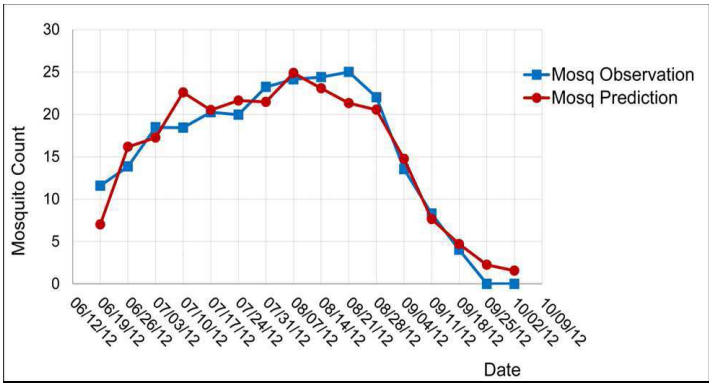

Figure 2.10 The observed and model-derived forecast of average mosquito counts per trap for Peel Region in 2012.

weekly on Laboratory of Mathematical Parallel Systems (LAMPS) (`http://www.lamps.yorku.ca/weeklyforecast2011`, `http://www.lamps.yorku.ca/weeklyforecast2012`) and a weekly report was sent to Peel Region public health department, Public Health of Ontario (PHO) and the Public Health Agency of Canada (PHAC). The forecasting results were shown in Figure 2.9 for both clusters and the total forecasting results were shown in Figure 2.10 respectively. These were mostly good agreement between forecast and observed mosquito abundance for both clusters separately (particularly

for cluster 1 in which mosquito abundance was highest), and combined for Peel Region.

2.3.4 Impact of Weather on the Distribution Property

Cluster Name	p Value for Coefficient of Temperature	p Value for Coefficient of Precipitation	Adjusted R-squared
Cluster 1	0.0112	0.2036	0.7863
Cluster 2	0.3984	0.3843	0.2385

Table 2.6 The P values and adjusted R^2 in the regression models for the two clusters.

Temperature, but not precipitation, was associated with significant variations in deviation D for data from traps of cluster 1, but neither were associated with significant variations in D for data from traps of cluster 2 (Table 2.6). This result indicated that the distribution deviation of cluster 1 has a strong negative linear relationship with mosquito seasonal daily average temperature. When the mosquito season daily average temperature decreases by 1 degree, D will increase by 0.02, which implies that the discrepancy between the theoretical and empirical gamma distribution increases as temperatures cool.

During the study period of the year 2004 to 2011, the distribution of mosquito abundance in 2009 deviated from the gamma distribution and in the other years, it followed a gamma distribution. There must exist a threshold which could separate

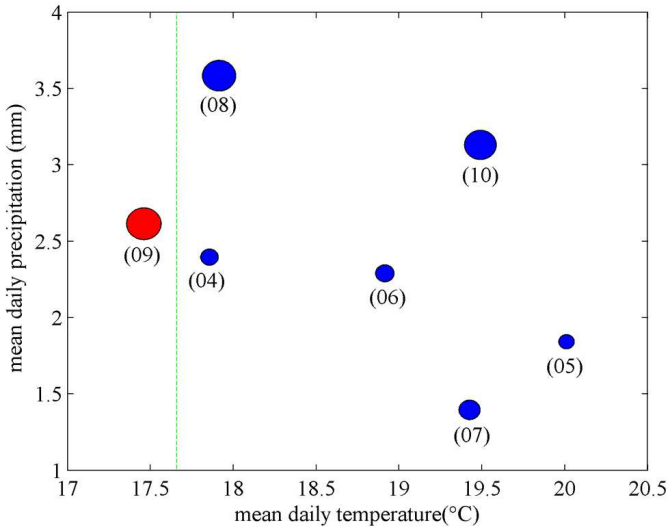

Figure 2.11 Mean temperature and precipitation in Peel Region from the year 2004-2011 and their relation to gamma-distributed mosquito abundance data in cluster 1.

the year in which the distribution of mosquito abundance followed a gamma distribution or not. For cluster 1, the lowest summer mean temperature during year 2004 to 2011 (17.46^oC) occurred in 2009, i.e. the year in which the mosquito abundance did not follow a gamma distribution. The year of 2004 had the lowest summer mean temperature of 17.86^oC among the years in which follow gamma distributions. Therefore the mean of 17.66^oC for these values could indicated a threshold above which mosquito abundance in cluster 1 follows a gamma distribution, and below which mosquito abundance deviates from a gamma distribution (indicated by a dashed line in Figure 2.11: The size of the dots was proportional to the average mosquito abundance of the year in the two clusters. Blue dots indicated mosquito density data

which followed gamma distributions, while the red dots showed mosquito density data which did not follow a gamma distribution. The green dashed line was the threshold which separated the years in which the mosquito counts followed gamma distribution and the year in which the mosquito counts deviated from a gamma distribution.).

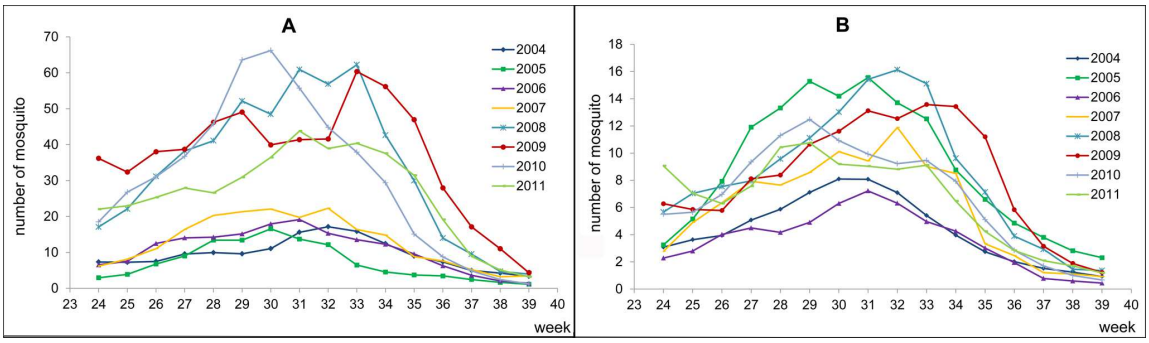

Figure 2.12 The time series of the weekly mean mosquito abundance for the two clusters in different years, A) Cluster 1; B) Cluster 2.

In most years from 2004 to 2011, the weekly average mosquito counts increased from the beginning of the surveillance season to a single peak at around week 30 to 33 for cluster 1 and week 29 to 33 for cluster 2, and then declined until the end of surveillance season (Figure 2.12). However, for cluster 1 in 2009, when the data diverged significantly from a gamma distribution, two peaks occurred, one in week 29 and the other in week 33. This could be consistent with unusually low temperatures reducing mosquito- breeding, development and activity in mid-summer (Figure 2.13).

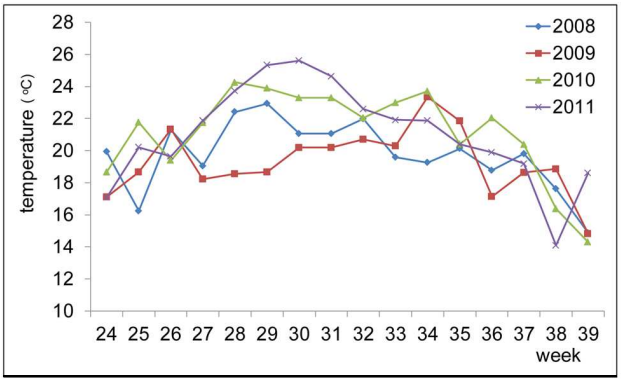

Figure 2.13 Weekly mean temperature from the year 2008 to 2011.

2.4 Discussion and Conclusion

The distributional properties of the mosquito abundance data from the year 2004 to 2011 in Peel Region were explored in this chapter. The clustering method was shown to be an effective method to capture the influence of geographic features on mosquito abundance in a region. Since the trap locations are permanently fixed in Peel Region from 2004 to current, the differences in physical, non-climatic environment characteristics of the sites where the traps are set should remain constant over time and therefore will not affect the cluster identification. But the composition of the clusters should be checked from time to time to keep the clustering method operationalised, as landuse changes could well affect their composition and this has not been included in this study. In each cluster, the mosquito abundance data followed a gamma distribution for most years except for cluster 1 in 2009. *Culex* mosquitoes are urban habitat mosquitoes, and their common breeding sites include roadside catch

basins, ditches, discarded tires, containers left outdoors and, in some circumstances, unused swimming pools (PPH 2008). Since female *Culex* mosquitoes require small pools of standing water to lay their eggs and have a preference for artificial breeding sites, they are found in greater numbers in mature suburban areas where open ditches and culverts provide favorable habitats for the mosquitoes to lay their eggs, resulting in high abundances. In contrast, areas such as recent housing developments without ditches or culverts may be less suitable for *Culex* mosquito and results in lower mosquito abundance. The analysis of associations of clusters with different landuse variables supported the geographic basis of the clusters despite a very small sample size to assess associations with individual landscape classification variables. The uncultivated meadow landclass (termed cultural meadow), associated with traps in cluster 1 comprises agricultural land that is no longer cultivated and is in transition to woodland, as is found in areas surrounding suburban developments (Barker et al. 2009). In contrast, woodland was associated with traps in cluster 2, and indeed woodland in the region predominates in areas distant from main housing areas (Figure 2.3). This finding is consistent with the observed capacity for mosquitoes in the vicinity of traps of cluster 1 to increase in numbers more markedly than those of cluster 2 when weather conditions for breeding improve. In our previous forecasting model studies, particularly high mosquito abundance, predictable by temperature

and rainfall, occurred in 2008 and 2009. Here we showed that the response in those years was in large part due to responses of traps in cluster 1, but not in cluster 2 (Figure 2.2).

The test results during both training and forecasting period of the statistical predictive models by clusters showed that forecasting accuracy was good. This proved that the forecasting by clusters is an effective method to increase forecast precision. Based on the number of probable and confirmed human cases, the WNv outbreak in 2012 was the worst ever in Peel Region since the virus was first detected in 2001. It is likely that weather played a key role, with an extremely mild winter followed by an early spring (EC 2012) likely enhancing overwinter survival of mosquitoes, then a long period of hot and humid conditions (EC 2012) that likely created near-perfect breeding environments for *Culex* species. Accounting for the clusters, our statistical model well forecasted mosquito abundance in 2012, although we caution that, due to the accuracy of the long-term weather forecast, the prediction results deteriorate for forwarding forecasting of greater than one week.

The multiple-linear-regression results suggested that the mean summer temperature (but not precipitation) had a strong negative relationship with the deviation from a gamma distribution for mosquito capture data from traps of cluster 1. This was not so for traps of cluster 2. Our observations suggested that a seasonal daily

average temperature of 17.66^oC could be a threshold temperature below which the mosquito abundance data deviate from a gamma distribution, and this was associated with a double peak in mosquito abundance. Wang et al. (2011) suggested that the unusually high peak mosquito values could be the result of either unusually wet and normally warm weather, or unusually hot and dry weather by the analysis of mosquito peak values each year. This indicate the threshold that separate the normal and abnormal weather conditions maybe the weather pattern, not only the mean summer temperature. Further study of this phenomenon, to more mechanistically incorporate it in forecasting models is needed.

The mosquito life cycle, reproduction rate, number of blood meals and breeding season are all closely related to climate and other environmental factors (Reeves et al. 1994). Experiences of the WNv outbreaks in the Canadian Prairies in 2007 (Artsob et al. 2009), and in Texas (Murray et al. 2013) and Ontario in 2012 (MOHLTC 2013) underlined the likelihood that increasing temperature and more abnormal weather events will drive outbreaks of mosquito-borne diseases. In turn this increases the need for effective forecasting to rationally target effective control of VBDs (NRC 2007). While the mechanisms linking climate and weather patterns to ecosystems such as WNv transmission cycles are very complex, here we have begun to develop methods of integrating weather conditions with geographical variations in landscape,

and our study identified methods for doing so. The clustering method we have developed allows trap data to be placed into separate clusters that correspond to habitat types that respond differently to variations in weather. The method may preclude the need for detailed landscape/habitat data, which are often unavailable or out of date, in correctly calibrating weather-based forecasting models from trap data. However, the method identified that accounting for geographic variations in habitat may considerably improve weather-based forecasting.

The study identified threshold climate conditions for responses of mosquito populations (in this case a mean summer 17.66^oC) to changes in weather so that we can better model and predict changes in mosquito abundance. Pragmatically we identified mean summer temperature conditions under which our forecasting models need modification to be more accurate.

3 Forecasting WNv activity in Greater Toronto Area under weather conditions by model selection

3.1 Introduction

It is already commonly recognized that the variations in weather greatly affects the abundance of the vectors including mosquito and the transmission of vector-borne diseases (VBD), such as West-Nile virus (WNv) and Lyme disease. It is anticipated that the increasing temperature and more frequent extreme weather events will increase the burden of predicting and effective control of the VBD (Health Canada 2008). To develop effective control strategies, various statistical models were built to forecast the mosquito population, WNv risk, and human incidence. Mosquito abundance and WNv transmission are both affected by weather and environmental factors, with temperature and precipitation considered key variables. The transmission of WNv is more strongly linked to the mosquito population (Turell et al. 2005)

and temperature (Diuk-Wasser et al. 2006, Reisen et al. 2008, 2006, Ruiz et al. 2010), while the impact of precipitation on WNv is complicated and comparatively weak (Epstein 2001, Sutherst 2004). A national study (Landesman et al. 2007) in the Unite State found inconsistent correlations between spatial patterns of precipitation and WNv incidence, which indicated that both the strength and the direction (positive or negative) of the effect of precipitation depend on the geographic area and the time period examined. There have been statistical modeling studies attempted to predict how climate change might affect the distribution of mosquito-borne diseases.

The recent work of Wang et al. (2011) discovered that weekly arithmetic means of mosquito counts of all traps in Peel region, Ontario follow a gamma distribution. A predictive statistical model for mosquito populations based on weather conditions was developed and optimism was provided for the development of weather-generated forecasting for WNv risk. Descloux et al. (2012) developed a climate-based multivariate non-linear statistical model using Support Vector Machines (SVM) technique to estimate the yearly risk of dengue outbreak in Noumena. Costa et al. (2015) studied the relationship between egg number and climate and environmental variables through Bayesian zero-inflated spatial-temporal models. An and Rocklöv (2014) associated dengue fever in Hanoi with the meteorological determinants by negative binomial model. In this chapter, we expanded the methods developed by Wang et al.

(2011) to establish forecasting models to predict the mosquito abundance in Greater Toronto Area (GTA), Ontario. Only weather conditions were included inWang et al. (2011)s model, more variables contributing to mosquito abundance would be explored. We also established new models to predict WNv risk and human incidence. The patterns of WNv risk and human cases are very complex with lots of zeroes in the data. Their patterns have not been fully investigated. Model selection has been used to select the predictive models for VBD (An and Rocklöv 2014, Marcantonio et al. 2015), but has not been used to determine the model itself. In this chapter, we proposed multiple models, such as Zero-Inflated Poisson (ZIP), gamma, Poisson, negative binomial, Zero-Inflated negative binomial distributions to predict the WNv risk and human incidence and employed model selection to choose the best fit models. The objective of this study is to develop accurate temporal models to forecast WNv vector mosquito population, WNv risk and human incidence using GTA mosquito surveillance data under weather changes. The models developed would contribute to build up the integrated real-time Early Warning and Response System (EWARS) for WNv in Ontario, which also helps for control and prevention of other mosquito-borne diseases.

3.2 Materials and Methods

3.2.1 Study site and program

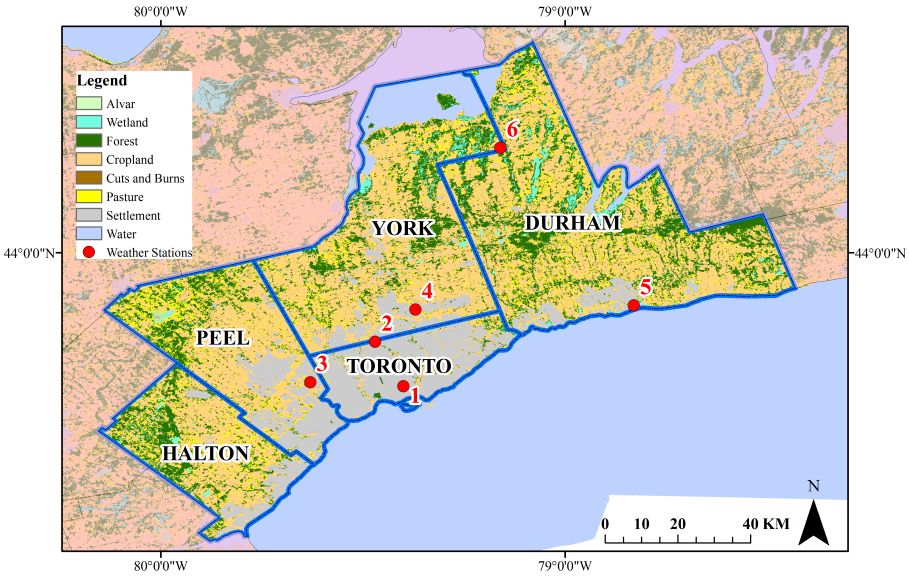

Figure 3.1 GTA landuse map and location of weather stations.
1-Toronto City, 2-Toronto North York, 3-Toronto Lester B. Pearson INTL Airport, 4-Toronto Buttonville Airport, 5-Oshawa WPCP, 6-Dora

GTA is located in southern Ontario, Canada (see Figure 3.1). Although often perceived as an urban area, 66% of the approximately 7,512 km^2 of land in the GTA is rural (Walton 2010). The GTA includes the city of Toronto and four regional municipalities (Durham, Halton, Peel and York) with a combined population of about 6.6 million (Statistic Canada 2012), making it the fifth largest populated area in North America and the largest one in Canada. Four of the regions (Halton, Peel, Toronto and Durham) of GTA lie along a continuous urbanized lakeshore and

shares prime access to the Great Lakes. Southern parts of the four regions are densely populated, and also concentrated with industry and commercial activities. Conversely, the northern parts consist more of agricultural industry. Although not lying along the lakeshore, York Region shares the similar structure. Because of this, we averaged the surveillance data from the five regions and used it to build the forecasting models to predict the WNv activity in GTA.

Climate of GTA has some fairly unique features, including four well-marked seasons. Winters are normally cool to cold and summers are warm to sometimes hot; monthly precipitation amounts average around 65 mm of rain or rain and snow equivalent, which makes it one of the most reliable precipitation regimes in the world, with markedly dry or wet spell both uncommon. Being situated in the midlattitudes, 43 degrees north, GTAs climate is influenced by the moving boundary between continental polar air that originates in northern Canada and maritime tropical air which forms over the Gulf of Mexico and subtropical North Atlantic (Gough 2000). At the local scale, Lake Ontario and the other Great Lakes have a modifying influence on GTAs climate, the lake effect (Auld and Service 1990).

3.2.2 Data collection and processing

Mosquito surveillance program in Ontario started in 2001 by the Ministry of Health and Long-Term Care (MOHLTC). The 36 Public health units in Ontario conduct mosquito surveillance weekly from June to October each year. Centers for Disease control (CDC) light traps have been prevailingly used to capture adult female mosquitoes. Traps are set up once a week and mosquitoes are collected, identified up to species level and counted. A real Time Reverse Transcriptase Polymerase Chain Reaction test is done to determine the WNv status of the various mosquito pools (PHO 2012). Since *Culex.pipens* and *Culex.restuans* are the mosquito species that are responsible for most WNv transmission (PHO 2012), it is reasonable to use the abundance data of *Culex.pipens/restuans* as mosquito count. The mosquito data of the five regions in GTA we used in this study were collected via this program. The same data smoothing technique used in Section 2.2.2 was applied to the mosquito data.

There are different ways to evaluate the prevalence of WNv transmission intensity in an area. The two most commonly used ones are minimum infection rate (MIR) and maximum likelihood estimation (MLE) (Gu et al. 2003). MIR is based on the assumption that infection rates are generally low and that only one mosquito is positive in a positive pool. MLE does not require the assumption of one positive

mosquito per positive pool, and provides a more accurate estimate when infection rates are high. In general, MIR and MLE are similar when infection rates are low. Both MIR and MLE can provide a useful, quantitative basis for comparison, allowing evaluation of changes in infection rate over time. In this study, we used MIR as an indicator of the WNv transmission intensity. MIR is the number of positive batches of infected mosquitoes of a given vector species divided by the total number of mosquitoes of a given vector species that are tested for the presence of the virus, expressed per 1000. Since we focused on WNv risk, the following *Culex* mosquitoes were employed in the measurement: *Culex pipiens*, *Culex pipiens/restuans*, and *Culex restuans*. That is, the formula for calculation of MIR for a given species is:

$$MIR = \frac{No\ of\ positive\ batches}{No\ of\ mosquitoes\ been\ tested\ for\ virus} \times 1000. \tag{3.1}$$

The weekly MIR of GTA was calculated from the datasheet provided by the Ontario mosquito surveillance program. The smoothing technique described in Section 2.2.2 was applied to the calculated MIR data too.

A human case is identified when a person visits a physician and the symptoms of WNv infection are detected. The health care provider then submits a blood samples to the MOHLTC Central Public Health Laboratory. If the laboratory tests are positive, the laboratory notifies both the local public health unit and the Public Health Division of MOHLTC (PHO 2012). In this study, the summation of the

human cases in the five health units of GTA from 2002 to 2012 was used to build the predictive models.

The weather data used as predictive variables was obtained from Canadas National Climate Archive (`http://www.climate.weatheroffice.gc.ca`). Because of the comparatively small scale of GTA, we neglected the temperature difference from south of lakeshore to north, and chose six weather stations (see Figure 3.1) at different locations which had complete record during the study period from 2002 to 2012. The average temperature and precipitation of the six weather stations were used for analysis. *ddm* and *ppm* defined in Section 2.2.5 were used to show the accumulated effect of temperature and precipitation.

Temperature and precipitation are crucial to mosquito abundance (Madder et al. 1983, Reisen et al. 2008, Wang et al. 2011). Mosquitoes are unable to regulate their body temperatures, thus depending on the surrounding temperature for warmth and growth (Shelton 1973). There are developmental temperature thresholds for the larval and pupal of mosquito to grow. Under certain temperature, the mosquitoes can not develop. The survival of the mosquito lava and pupal are affected by temperature too. Shelton (1973) found that it causes 100% mortality for larvae and pupae at temperature 32^oC and 35^oC. Therefore an optimal temperature must exist for mosquito lava and pupa to develop. The reaction of mosquito abundance to precipitation is

similar to what it has to temperature (Bolling et al. 2005, Madder et al. 1983, Pecoraro et al. 2007). The immature life stages of mosquito prefer water habitats with a high organic content (Turell et al. 2005). The standing water formed after rainfall provides habitat for lava and pupa to develop. If there is too much rainfall, it may flush away the mosquito eggs, lavas and pupas, causing the developmental rate of mosquito to decline. There should be an optimal precipitation amount for mosquito to develop. The effect of temperature and precipitation on WNv transmission and human cases are similar to what they have on mosquito (Reisen et al. 2006, Ruiz et al. 2010). The observation from the study of Reisen (1995) showed the mortality rate of mosquito against temperature was a U-shape function. Wimberly et al. (2008) used second-order trend surface of temperature and precipitation to analyse the relationship between WNv incidence and the environmental divers. In this study, we employed quadratic form of temperature and precipitation to develop the predictive models for *Culex* mosquito abundance, MIR and human infection incidence.

3.2.3 Modeling the impact of weather on *Culex* mosquito abundance

Predictive variables for *Culex* mosquito abundance model

The quadratic form of ddm_k and ppm_l were used as predictive variables to show the accumulated effects of temperature and precipitation. We used ddm_k and ppm_l

for k and $l = 1, \cdots, 60$ to run our models. The model selection criterion was applied to choose the most significant ddm_k and ppm_l as predictive covariates. The time-series of *Culex* mosquito counts indicated the *Culex* mosquito abundance of the previous weeks have impact on the current week's abundance (see Figure 3.2). Accordingly we included the first order autoregressive (AR) term as potential explanatory variables.

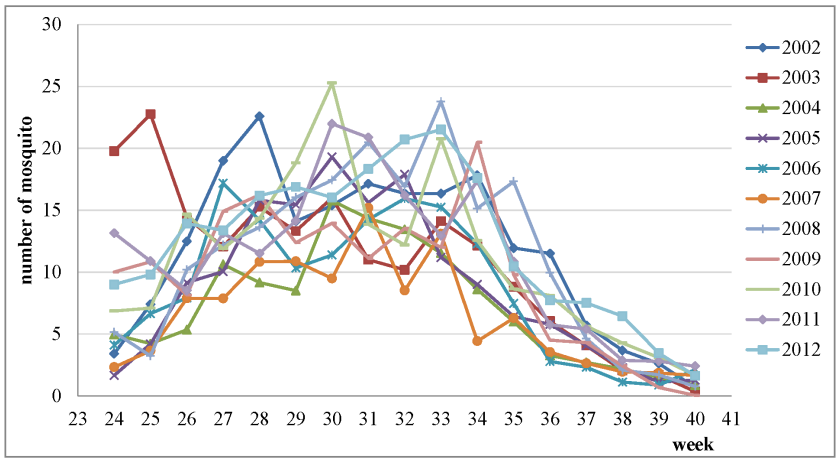

Figure 3.2 The time series of the weekly mean Culex abundance in GTA in different years.

Predictive model for *Culex* mosquito abundance

The *Culex* mosquito abundance model is same as the one for Peel Region built in Section 2.2.5 except the predictive variables. Fewer vicariates were used to capture the main effects of the weather conditions on mosquito abundance rather than the

predictive accuracy. The predictive models could be written as:

$$\begin{aligned}\log(\mu) = \quad & \alpha_0 + \alpha_1 ddm_k + \alpha_2 ppm_l + \alpha_3 ddm_k \times ppm_l + \alpha_4 ddm_k^2 + \alpha_5 ppm_l^2 + \\ & \alpha_6 AR \times I_{(t>1)} + \alpha_7 I_{t=1)},\end{aligned} \tag{3.2}$$

where μ is the weekly mean mosquito abundance, AR is the first order auto regression term of mosquito abundance, $I(t = 1)$ is an indicator variable to show the first week mosquito abundance has no AR term, meanwhile $I(t = 1)$ shows the AR term for mosquito abundance exists from the second week. αs are the regression coefficients.

Two models were considered in this study. The first model used the reciprocal function as link function and the second model applied log function as link function. Cross validation was used as model selection criterion. Among the data from the year 2002 to 2012, the data was partitioned with 10-year data as training data and the remaining one year as testing data. The cross validation process was then repeated 11 times so that the data of each year was used exactly once as the validation data. The best fit model was determined by the root-mean-square error (RMSE) (Hu et al. 2006, Makridakis et al. 2008), which is defined as:

$$RMSE = \sqrt{\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2/mn}, \tag{3.3}$$

where y_{ij} is the predicted value of the ith study year in the jth week and $\hat{y}_{ij}$ is the corresponding observation. The smaller the RMSE is the better the model in terms

of goodness of fit. The BIC was also used to further validate the model, which was written as:

$$BIC = -2 \times \log(L) + k \times \log(n), \tag{3.4}$$

where n is the number of observations, L is the maximized value of the likelihood function for the estimated model, and k is the number of parameters. The model with the smallest BIC was chosen to be the best fit model.

3.2.4 Modeling the impact of weather on WNv risk

Predictive variables for WNv risk model

Similar to *Culex* mosquito abundance predictive model, quadratic form of ddm_k and ppm_l were used as predictive variables to show the accumulated effects of temperature and precipitation on WNv risk. The ddm_k and ppm_l for k and l from 1 to 80 days prior to the surveillance date were explored as the explanatory variables. Ruiz et al. (2010) had discovered MIR was a first order autoregressive process in northeast Illinois, USA, therefore the first order AR term was included in the predictive model. In order to investigate the effect of *Culex* mosquito count on the WNv transmission (Barker et al. 2009), the *Culex* mosquito abundance data from 1 up to 4 weeks before the capture data were considered to be predictive variable.

Predictive model for WNv risk

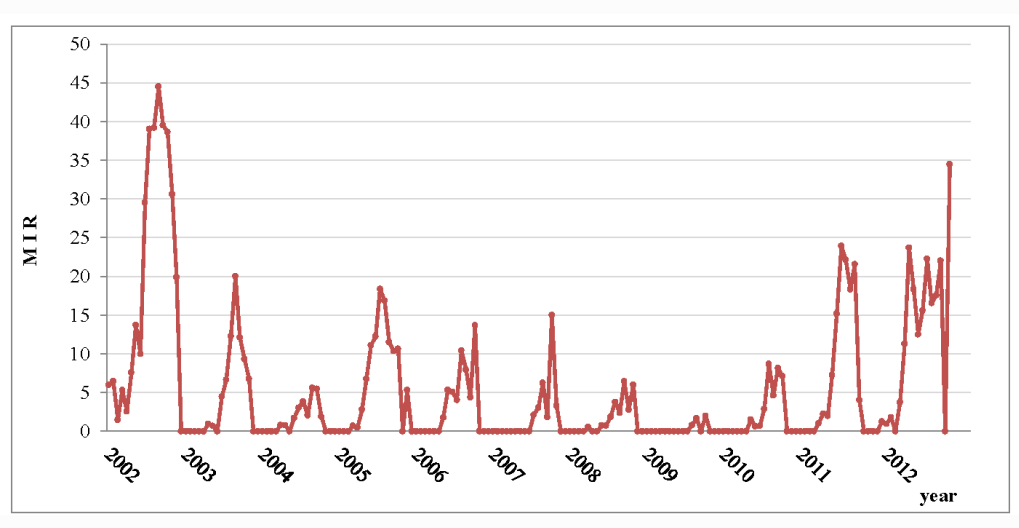

Figure 3.3 The MIR time series plot of GTA from the year 2002 to 2012.

The weekly time series plot of MIR from the year 2002 to 2012 (Figure 3.3) has shown a very complex pattern with some zero counts. No researchers have studied the distribution pattern of MIR. Zero-Inflated Count Models have been proposed for the situations where the data generating process results into too many zeroes. Their applications include many areas of interest such as public health, epidemiology, sociology, psychology, and engineering. Lambert (1992) described the zero-inflated Poisson (ZIP) regression models with an application to defects in manufacturing; Lee et al. (2001) proposed ZIP to model manual handling injuries data. Martin et al. (2005) proposed a framework for using zero-inflated models to describe the ecology presence/absence and count data. Costa et al. (2015) studied the relationship between egg number and climate and environmental variables through Bayesian zero-

Model	Probability Distribution Function	GLM Model	Parameters
gamma	$f(y;\alpha,\beta)=\frac{y^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$	$\log(\mu)=X\eta$	y is the random variable, α and β are shape and scale parameters, μ is the mean of random variable, X is the predictive variable vector, η is the coefficient vector.
poisson	$f(y;\lambda)=\Pr(Y=y)=\frac{\lambda^{y}e^{-\lambda}}{y!}$	$\log(\mu)=X\eta$	y is the random variable, λ is the mean parameter, μ is the mean of random variable, X is the predictive variable vector, η is the coefficient vector.
negative binomial	$f(y;r,p)=\Pr(Y=y)=\binom{y+r-1}{y}(1-p)^{y}p^{y}$	$\log(\mu)=X\eta$	y is the random variable, $r>0$ is the number of failures until the experiment is stopped, $0<p<1$ is the success probability in each experiment, X is the predictive variable vector, η is the coefficient vector.
zero-Inflated poisson	$P(Y=y)=\begin{cases}\omega+(1-\omega)\exp(-\lambda) & y=0,\\ (1-\omega)\frac{\exp(-\lambda)\lambda^{y}}{y!} & y>0\end{cases}$	$\begin{cases}\log it(P(\mu=0))=X\eta_1 & \mu=0\\ \log(\mu)=X\eta_2 & \mu>0\end{cases}$	y is the random variable, λ is the mean parameter, ω is measure of the extra proportion of zero in a sample unit, $0<\omega<1$, μ is the mean of random variable, X is the predictive variable vector, η is the coefficient vector.
zero-Inflated negtive binomial	$\Pr(Y=y)=\begin{cases}\pi+(1-\pi)\frac{\Gamma(\theta)}{\Gamma(a)}(\frac{\theta}{\theta+\mu})^{\theta} & y=0\\ (1-\pi)\frac{\Gamma(\theta+y)}{\Gamma(a)\Gamma(y+1)}(\frac{\theta}{\theta+\mu})^{\theta}(\frac{\mu}{\theta+\mu})^{y} & y>0\end{cases}$	$\begin{cases}\log it(P(\mu=0))=X\eta_1 & \mu=0\\ \log(\mu)=X\eta_2 & \mu>0\end{cases}$	y is the random variable, μ and θ are the mean and the size parameters, respectively, π is measure of the extra proportion of zero in a sample unit, $0<\pi<1$, μ is the mean of random variable, X is the predictive variable vector, η is the coefficient vector.

Table 3.1 The predictive models.

inflated spatial-temporal models. In this study, we developed several models such as gamma, Poisson, negative binomial, Zero-Inflated negative binomial and ZIP (See Table 3.1) to fit the data. Leave-one-out cross validation, RMS error and BIC, same as defined in the mosquito predictive model in Section 3.2.3, were used as the model selection criteria.

Predictive model for WNv human incidence

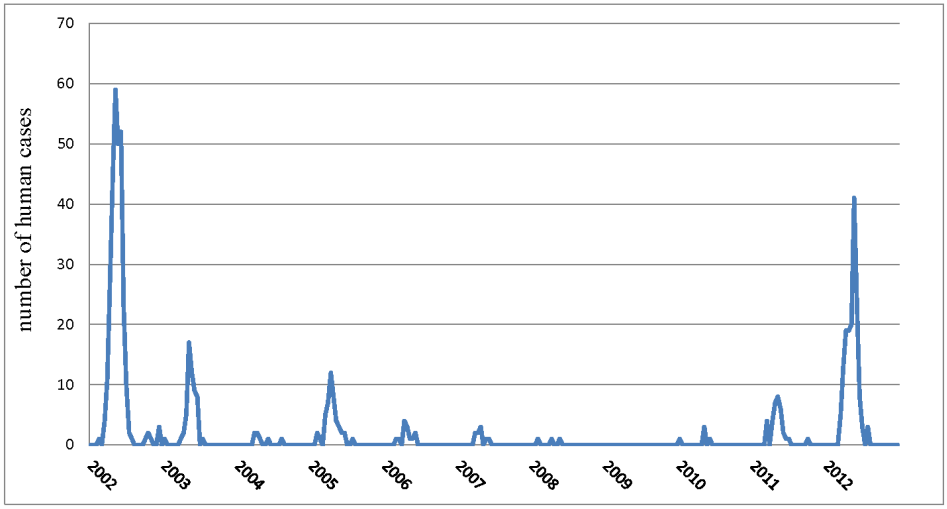

Figure 3.4 The WNv human cases time series plot of GTA from the year 2002 to 2012.

The time series plot of WNv human cases in GTA from the year 2002 to 2012 (Figure 3.4) showed a more complex pattern than that of MIR with a lot of zeroes. In order to access the forecasting accuracy of different methods, the following models were tested: 1) Gamma model 2) Poisson model 3) negative binomial model 4) Zero-Inflated Poisson model 5) Zero-Inflated negative binomial model. The accuracy of each model was determined by RMSE and BIC through leave-one-out cross validation, same as defined in the mosquito predictive model.

3.3 Results

3.3.1 The Most Significant Temperature and Precipitation Conditions for *Culex* mosquito abundance and the best fit model

Model	Link Function	RMS error	BIC
1	reciprocal	16.9102	923.21
2	Log	16.0381	845.76

Table 3.2 The RMS error and BIC of mosquito predicting models.

Coefficient	Estimated Value	Standard Error	P Value
α_0	-2.3686	0.5244	$1.14\times e^{-5}$
α_1	0.3659	0.0593	$4.54\times e^{-9}$
α_2	0.8712	0.2425	0.0004
α_3	-0.0398	0.0127	0.0022
α_4	-0.0089	0.0021	$3.46\times e^{-5}$
α_5	-0.0713	0.0316	0.025
α_6	0.0954	0.0068	$<2\times e^{-16}$
α_7	1.0156	0.1234	$3.71\times e^{-14}$

Table 3.3 The regression parameters of the mosquito predicting model.

The two models (reciprocal and log link functions) were constructed with ddm_k and ppm_l for k and $l = 1, \cdots, 60$ prior to the surveillance date (i.e., $60 \times 60 = 3600$ computations). The RMSE and BIC of each model were compared in Table 3.2. Between the two models, model 2 (log link function) achieved smaller RMSE of 16.0381 and BIC of 845.76 at the same time. Therefore model 2 was chosen to be the

best-fit model to predict *Culex* in GTA and the corresponding regression coefficients, standard error and p-values were listed in Table 3.3. All the predictive variables were significant, which confirmed that the temperature and precipitation have a big effect on the *Culex* mosquito abundance and the *Culex* mosquito abundance is a first-order autoregressive process. The contour map of the RMS of model 2 was showed in Figure 3.5 with different combination of ddm_k and ppm_l as predictive variables. The combination of ddm_{10} and ppm_{57} has achieved the smallest RMSE, therefore been chosen to be the most significant predictive variables.

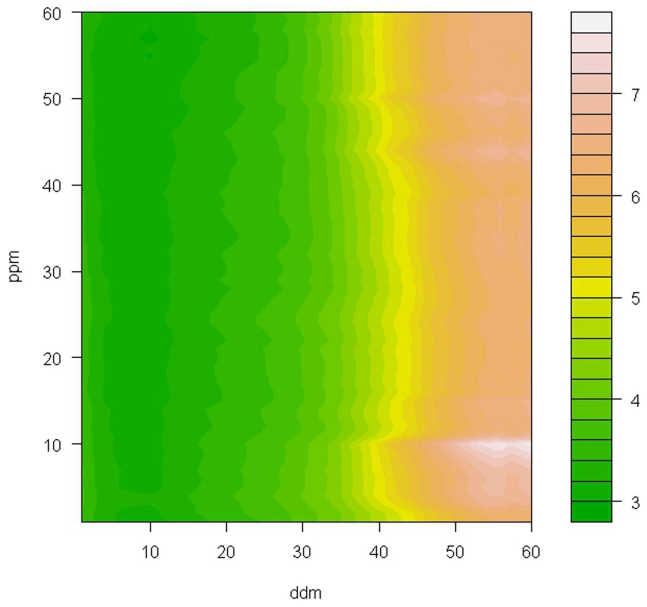

Figure 3.5 The mosquito forecasting model RMSE using temperature (ddm) and precipitation (ppm) as covariates.

By leave-one-out cross validation, the results of the predicted data versus the observed data from the year 2002 to 2012 were showed in Figure 3.6. The model

could predict the *Culex* mosquito abundance trends but the prediction of the peak times had about one week delay. The model fitted very well with the observation values for most of the years except for the year 2011 and 2012. The model predicted a higher peak than the observations in 2011. In 2012, the predicted values in the later half mosquito surveillance season were higher than observations.

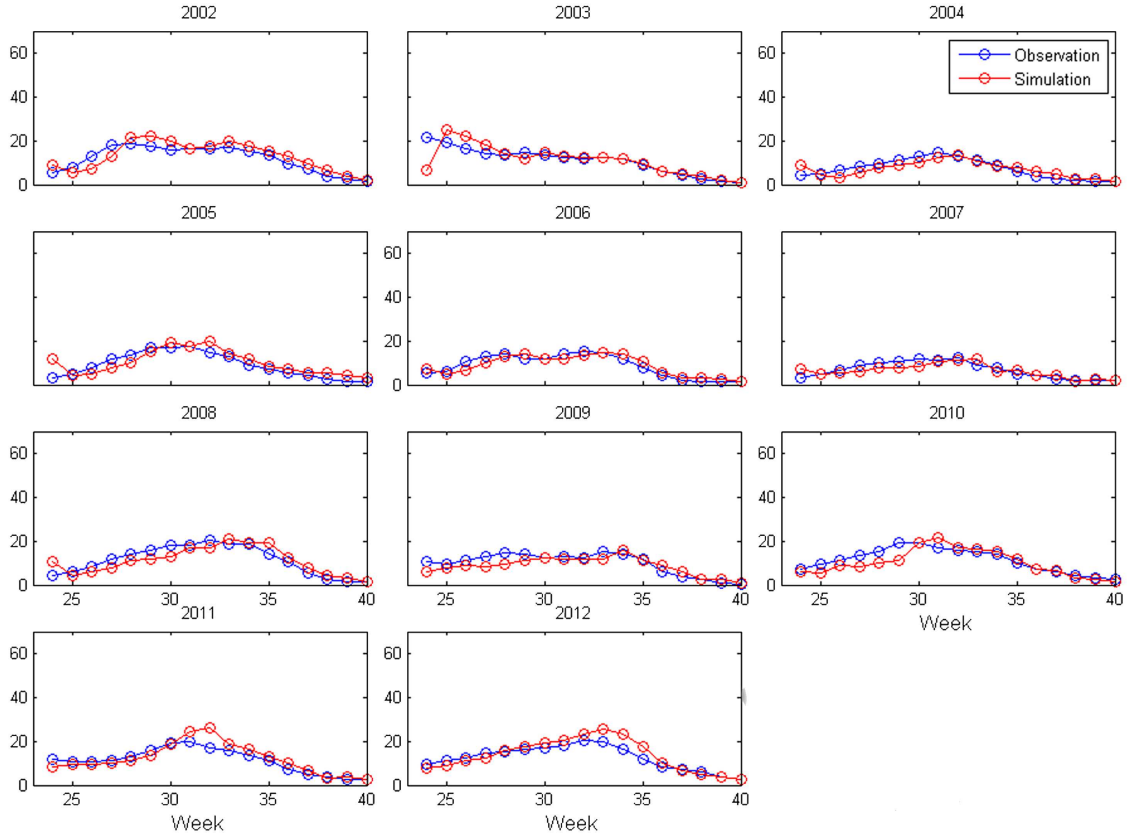

Figure 3.6 The observation versus simulation of mosquito counts in GTA from 2002-2012.

Model	Distribution Function	RMS error	BIC
1	gamma	47.1813	647.57
2	Poisson	6.9326	963.64
3	negative binomial	NA	NA
4	zero-Inflated Poisson	2.4394	915.19
5	zero-Inflated negative binomial	NA	NA

Table 3.4 The RMSE and BIC of WNv risk models.

3.3.2 Result of the WNv Risk Predictive Model

Five models (gamma, Poisson, negative binomial, Zero-Inflated negative binomial and Zero-Inflated Poisson) were constructed with ddm_k and ppm_l for k and $l = 1, \cdots, 80$ prior to the surveillance date (i.e., $80 \times 80 = 6400$ computations). The candidate predictive variables included one time and quadratic terms of *ddm*, *ppm*, and their interactions, AR term, and *Culex* mosquito abundance data from 1 to 4 weeks prior to the capture date. For each of the five models, the predictive variables were selected by RMSE and verified by BIC, which were showed in Table 3.4. Among the five models, negative binomial and Zero-Inflated negative binomial did not work for MIR predictive model since the algorithms did not convergent. Model 4 (Zero-Inflated Poisson) achieved the smallest RMSE of 2.4394. The BIC of ZIP model (915.19) was bigger than the BIC of gamma model (647.57) since the number of parameters ZIP model need to estimate was twice as much as that of gamma model and BIC gives a bigger penalty for more variables in the models. The RMSE of

gamma model (47.1813) was much bigger than that of ZIP model. ZIP model was chosen to be the best-fit model to predict MIR in GTA and the predictive variables were *ddm* and *ppm*. The first order AR and the *Culex* mosquito count one to four weeks prior to the capture date could not contribute to decreasing the RMSE, and were not chosen to be the predictive variables.

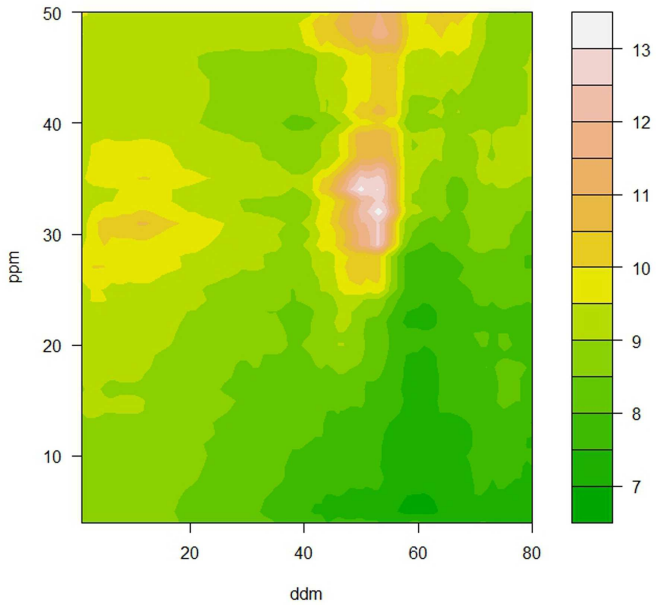

Figure 3.7 The MIR forecasting model RMSE using temperature (*ddm*) and precipitation (*ppm*) as covariates.

The predictive model built by ZIP described the following process:

$$\begin{cases} logit(P(\mu_{MIR} = 0)) = X\beta_1, & \mu_{MIR} = 0, \\ \log(\mu_{MIR}) = X\beta_2, & \mu_{MIR} > 0 \end{cases} \tag{3.5}$$

where μ_{MIR} is the weekly mean MIR, $P(\mu_{MIR} = 0)$ is the probability of observed extra zero value of MIR other than predicted by Poison distribution, X is the predictive

vector:

$$X = (1, ddm_k, ppm_l, ddm_k \times ppm_l, ddm_k^2, ppm_l^2) \tag{3.6}$$

and β_1, β_2 are the regression coefficients.

Coefficient	Estimated Value	Standard Error	P Value
β_{10}	2.9202	3.1427	0.3528
β_{11}	1.4199	0.7655	0.0636
β_{12}	0.4869	0.5262	0.3549
β_{13}	-0.0087	0.0514	0.8661
β_{14}	-0.1258	0.0476	0.0082
β_{15}	0.0362	0.023	0.115
β_{20}	3.6399	0.8693	$2.82\times e^{-5}$
β_{21}	-0.8583	0.1494	$9.23\times e^{-9}$
β_{22}	0.0902	0.1005	0.3695
β_{23}	-0.0028	0.0083	0.7327
β_{24}	0.0597	0.0066	$<2\times e^{-16}$
β_{25}	-0.007	0.0033	0.7327

Table 3.5 The regression parameters of the WNv risk predicting model.

The contour map of the RMSE of ZIP model was showed in Figure 3.7 with different combination of ddm_k and ppm_l as predictive variables. The combination of ddm_{61} and ppm_5 has achieved the smallest RMSE, therefore been chosen as the most significant predictive variables. The regression coefficients, standard error and p-values were listed in Table 3.5. The coefficient of ddm_{48} and ${ddm_{48}}^2$ were significant, which confirmed that the temperature had big impact on the WNv transmission and the impact of precipitation was not significant.

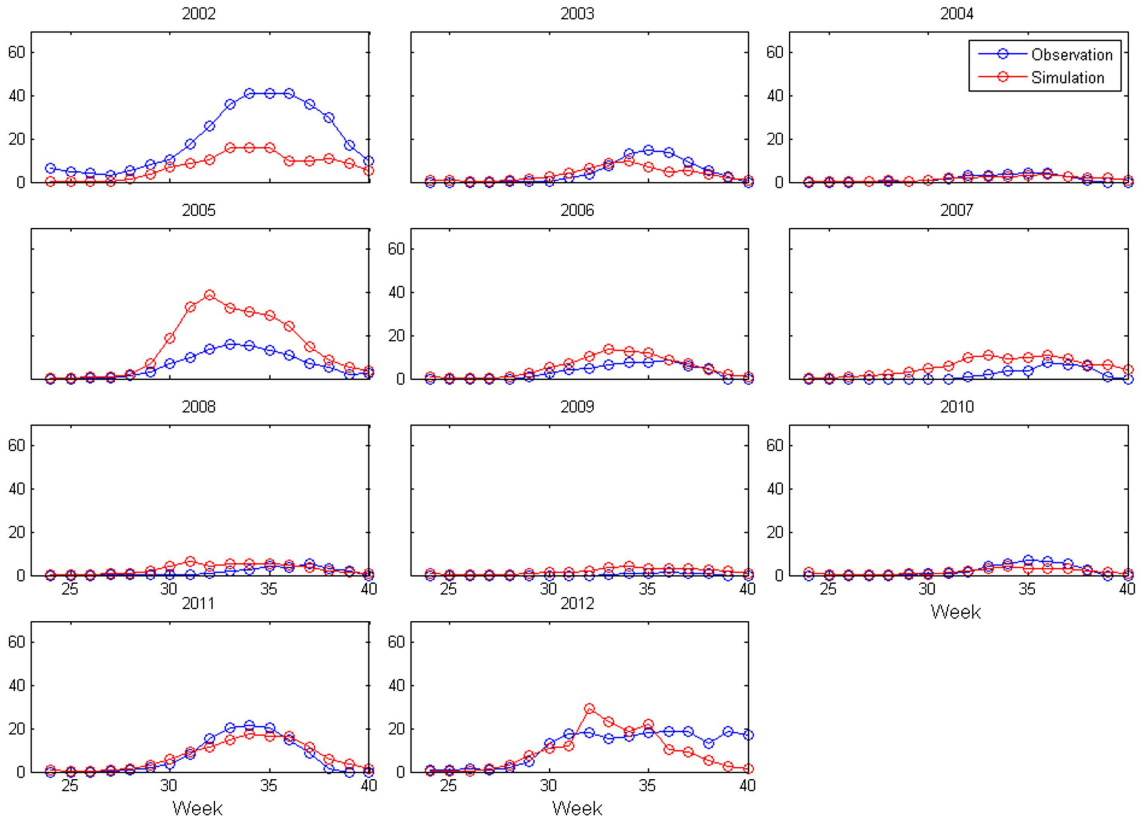

Figure 3.8 The observation versus simulation of WNv risk in GTA from 2002-2012.

By leave-one-out cross validation, the performance of the best-fit model from the year 2002 to 2012 was depicted in Figure 3.8. The simulation could capture the zeroes in the MIR counts and work very well on predicting the trend of WNv risk. The model overestimated MIR for the year 2005 and underestimated the peak value for the year 2002. In the year 2012, there was a big outbreak of WNv in North America. The MIR was continually high until the end of mosquito season. The model did not work well to forecast the high MIR value at the end of the mosquito

surveillance season.

3.3.3 Result of the Human Cases Predictive Model

Among the five models, the ZIP achieved the smallest RMSE (see Table 3.6) and was chosen to be the best fit model to predict the WNv human cases in GTA. The predictive variables included *ddm*, *ppm* and AR.

Model	Distribution Function	RMS error	BIC
1	gamma	299.49	263.56
2	Poisson	5.4321	556.6
3	negative binomial	NA	NA
4	zero-Inflated poisson	4.528	529.25
5	zero-Inflated negative binomial	NA	NA

Table 3.6 The RMSE and BIC of WNv human cases models.

The predictive model built by ZIP was as follow:

$$\begin{cases} logit(P(\mu_{hum}=0)) = X\gamma_1, & \mu_{hum}=0, \\ \log(\mu_{hum}) = X\gamma_2, & \mu_{hum}>0 \end{cases} \tag{3.7}$$

where μ_{hum} is the weekly mean human cases, $P(\mu_{hum}=0)$ is the probability of observed extra zero value of human cases other than predicted by Poison distribution, X is the predictive vector:

$$X = (1, ddm_k, ppm_l, ddm_k \times ppm_l, ddm_k^2, ppm_l^2, AR) \tag{3.8}$$

and γ_1, γ_2 are the regression coefficients(see Table 3.7 for the values of γ_1 and γ_2, the corresponding standard errors and P values).

Coefficient	Estimated Value	Standard Error	P Value
γ_{10}	-0.5693	8.2269	0.9448
γ_{11}	2.2797	1.4968	0.1278
γ_{12}	-6.0646	3.1155	0.0516
γ_{13}	0..4414	0.2618	0.0679
γ_{14}	-0.1825	0.0792	0.0213
γ_{15}	0.1751	0.2659	0.5101
γ_{16}	0.0439	0.0028	0.0838
γ_{20}	3.1971	2.8836	0.268
γ_{21}	-0.0792	0.426	0.852
γ_{22}	-3.5674	0.7774	$4.45\times e^{-6}$
γ_{23}	0..2910	0.0595	$0.87\times e^{-7}$
γ_{24}	-0.0057	0.0018	0.756
γ_{25}	-0.0092	0.0337	0.785
γ_{26}	0.0439	0.0028	$<2\times e^{-16}$

Table 3.7 The regression parameters of the WNv human cases predicting model.

The contour map of the RMSE of WNv human cases predictive model was showed in Figure 3.9 with a different combination of ddm_k and ppm_l as predictive variables. ddm_{67} and ppm_{43} had achieved the smallest RMSE and were chosen to be the predictive variables. The coefficients of $ddm_{67}, ddm_{67} \times ppm_{43}$ and AR were significant, which indicated the temperature and precipitation had a great impact on the WNv human incident and the WNv human incident was a first order AR. The MIR of previous weeks did not contribute to decreasing RMSE since the impact of MIR had already included in the temperature and precipitation. The performance of the best-fit model for human cases was depicted in Figure 3.10. The Zero-Inflated Poison

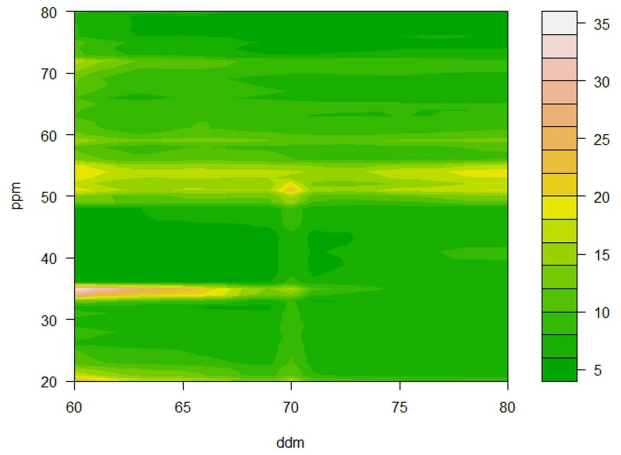

Figure 3.9 The WNv human cases forecasting model RMSE using temperature (ddm) and precipitation (ppm) as covariates.

generalized linear model worked very well to capture the extra zeroes in the human cases data. The predicted peaks was one week later than the observed ones and the predicted magnitude of the peak values were close to the observations, though it was under- or overestimated (see 2003, 2005 and 2012).

3.4 Discussion and conclusions

We developed forecasting models to predict the *Culex* mosquito abundance, the WNv risk and human incidence in GTA under weather changes. The weather conditions that affect the mosquito abundance and WNv transmission were examined and the most significant temperature and precipitation were given in each case. Different models were compared by using the surveillance data of *Culex* mosquito abundance,

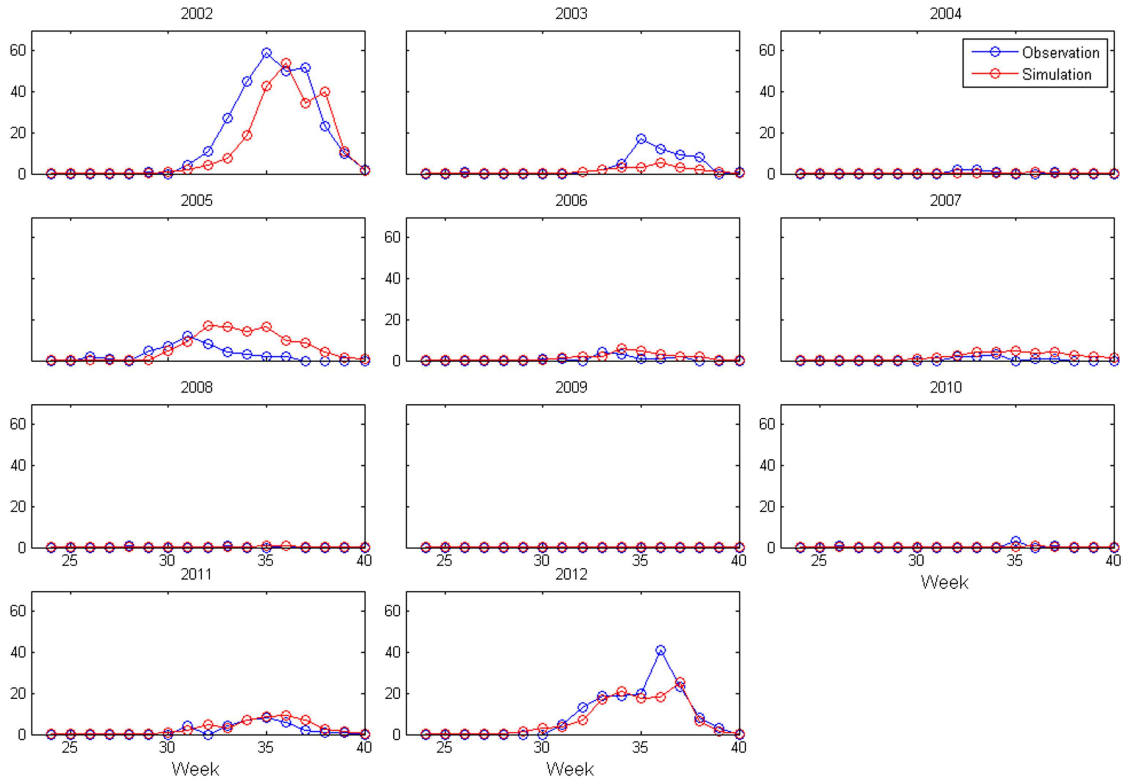

Figure 3.10 The observation versus simulation of WNv human Cases in GTA.

MIR, and human incidence from the year 2002 to 2012 and the best fit models were chosen by model selection method for each case. The predictions were in a good agreement with the observations for the period from 2002 to 2012. The model selection was proved to be an effective way to compare different models. These models chosen could be used by the public health authorities to forecast the WNv risk one week ahead. Since it still remains difficult to get accurate weather forecast over long period, the predicting results of more than one week ahead are not very reliable.

The accumulated degree day had a strong positive correlation with the *Culex* mosquito abundance, the WNv risk and the human incidence in GTA. This result agreed with previous studies (Reisen et al. 2006, Ruiz et al. 2010, Tibbetts 2007, Wang et al. 2011). Higher temperature will increase the mosquitos rate of reproduction and number of blood meals, prolongs their breading season, and shortens the maturation period for the microbes they disperse. The accumulated precipitation had a negative correlation with the *Culex* mosquito abundance, the WNv risk and the human incidence in GTA. This was because the dominant mosquito species which responsible for the transmission of WNv, the *Culex* species, are urban habitant mosquito. Their common breeding sites include roadside catch basins, ditches, discarded tires, containers left outdoors and, in some circumstances, unused swimming pools (Walsh et al. 2008). Since female *Culex* mosquitoes require small pools of standing water to lay their eggs and have a preference for artificial breeding sites, more precipitation will reduce lava survival through flushing effect (Koenraadt and Harrington 2008) and reduce trap counts (DeGaetano 2005) used for estimating mosquito populations. The time series of mosquito count, and MIR human cases were verified to be a first-order autoregressive process. The prediction accurate can be improved by including the first order autoregressive term in the models. The clustering technique is not applied for the forecasting models in GTA. The surveillance

data in GTA is varied in length and the normal distance-based clustering methods are not applicable. Since the covarites in the forecasting models include exogenous variables and past observations, the normal model-based clustering methods are not appropriate either. In next chapter, We developed a new model-based clustering method which can be applied to classify the above surveillance data.

4 Mixture Markov regression model for mosquito time series data

4.1 Introduction

Cluster analysis divides data into cohesive groups based on measured characteristics. It is a fundamental and broadly used methodology in data mining, image analysis, and gene expression data analysis and so forth. Clustering could be roughly classified as distance-based clustering (e.g., (Hartigan and Wong 1979)) and model-based clustering (e.g., (Banfield and Raftery 1993) and (Fraley and Raftery 2002)). In practice, we often face the situations where the clustering objects are in non-vector forms (such as time series and longitudinal data) or of different lengths. Special cases include clustering the gene expression data (Yeung et al. 2001), clustering and visualization of navigation patterns on a website (Cadez et al. 2000), identifying mid-latitude cyclones based on their temporal evolution (Blender et al. 1997) and

clustering different decay patterns of the market share data of movies (Wedel and Kamakura 2012). Since the standard distance based clustering is normally built on the assumption that the data can be represented as multivariate vectors of fixed dimensionality, it is not appropriate for this kind of clustering problems. As opposed to distance based clustering, the model-based clustering, particularly the finite mixture model approach, (Chamroukhi et al. 2011, Grün and Leisch 2008) offers a very natural alternative for the above problems. It can not only deal with the data of varying length but also incorporate the prior knowledge of data into the model, such as time series which depends on some covariances. In this chapter, we focus on the model-based clustering method of time series with Markov property. especially we will use a finite mixture of the regression model.

In the finite mixture approach, the data probability density function is assumed to be a combination of a finite number of K different components densities, each component density corresponds to a cluster. It is an extremely flexible method of modeling in statistics and the areas of application range from biology and medicine to physics, economics, and marketing. A comprehensive review of the finite mixture model is given in McLachlan and Peel (2004). Various approaches to estimate the parameters of the finite mixture models have been suggested over the years, from the moments method (Pearson 1894) to the graphical techniques(Harding 1949) and

nowadays the maximum likelihood(ML) approach (Basford and McLachlan 1985). The ML estimator has to be computed iteratively and is usually performed by the expectation-maximization(EM) algorithm (Dempster et al. 1977).

Model-based clustering can be naturally extended to analyse time series data. A finite mixture of Markov chains (Cadez et al. 2003, Poulsen 1990) has been widely used to cluster time series. While simple Markov chain is not good enough in some applications, hidden Markov model(HMM) (Rabiner 1989) can give more satisfied results due to their ability of model the versatility and capturing non-linear relationships. The application of HMM is rapidly increasing in speech recognation(Rabiner et al. 1989), image analysis (Eickeler et al. 2001) and bioinformatics(Schliep et al. 2003). Mixture of autoregressive moving average (ARMA) models and autoregressive integrated moving average (ARIMA) models have also been applied extensively for time series analysis (Liao 2005). Xiong and Yeung (2004) studied the clustering of data patterns that were represented as sequences or time series possibly of different lengths by using mixtures of ARMA models. If the time series can be modeled as a function of exogenous variables, mixture of the regression models is applicable. Besides normal mixture (DeSarbo and Cron 1988), a large number of mixture of generalized linear models(GLM) have been studied (see review of McLachlan and Peel (2004)) and a more general non-parametric form is given by Gaffney and Smyth

(1999).

Although these approaches work successfully most of the time at classifying time series, they all use only one global model (regression or autoregressive) within each cluster. As time series data is very unlikely to be independent, only regression model may not able to describe the data generating process well.

Regression methods have long been applied on time series analysis. Most of the work have been focused on linear regression of Gaussian time series (Anderson 1954, Fuller 2009), which cannot be directly applied on binary or counting data. For binary time series, Cox (1970) proposed Markov chain for an autoregressive logistic model, in which the linear predictor included both the covariates and also a finite number of past outcomes. Kalbfleisch and Lawless (1985) proposed to analyze categorical panel data by Markov model. The asymptotic theory of non-stationary categorical time series was established by Kaufmann (1987). The extension of generalized linear regression of time series was discussed by West et al. (1985). Fahrmeir (1989) proposed a generalized Kalman filter which can be applied on non-Gaussian time series. There were other models developed for non-Gaussian time series. For example, Azzalini (1982) proposed a Markov model to analyze time series with gamma distributed observations. Time series with a known exponential marginal density was discussed by Lawrance and Lewis (1985). However, these models were not formulated

in regression setting to include exogenous covariates.

Zeger and Qaqish (1988) proposed Markov regression models for time series using a quasi-likelihood approach. The model describes the mean response as a function of the covariates and past observations. These types of Markov models were referred by Cox et al. (1981) as being "observation-driven". The time dependence was modeled by the conditional expectation of the current observation on the past observations. These Markov models specify conditional distributions including Gaussian, Binomial, Poisson, Gamma, and other exponential family distributions. The first and second conditional moments were modeled explicitly as functions of covariates and past outcomes.

In this chapter, we propose to use Zeger and Qaqish (1988)'s quasi-likelihood approach to model time series. We provide a quasi-likelihood formulation for mixtures of time series. We develop a quasi-likelihood EM algorithm to deal with the missing data problem. The model parameters are estimated through mixture quasi-likelihood estimating equations. This chapter is organized as follow. Section 2 reviewed the finite mixture of GLM and the MLE via the EM algorithm. Section 3 focused on QL and the Markov regression models for time series discussed by Zeger and Qaqish (1988). In section 4, we presented a novel model-based clustering algorithm to cluster time series data which follows a finite mixture of GLM with Markov process.

Experimental results with simulated data were given in section 5. In section 6, the algorithm was applied to the mosquito surveillance data in Peel Region, Ontario. Section 7 concluded the paper with a summary and discussion.

4.2 mixture of GLM

4.2.1 Mixture of regression models

Consider dependent variable $\boldsymbol{y}$ with covariates of $\boldsymbol{x}$. The expectation and variance of $\boldsymbol{y}$ are denoted as $\boldsymbol{\mu}$ and σ^2 respectively. Let each component of $\boldsymbol{y}$ has a distribution in the exponential family , which can be written as

$$f_{\boldsymbol{y}}(y \mid \vartheta, \phi) = exp\{\frac{y\vartheta - b(\vartheta)}{a(\phi)} + c(y, \vartheta)\} \tag{4.1}$$

for some specific function $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ and parameters ϑ and ϕ. For the exponential family, it can be derived that

$$E(\boldsymbol{y}) = \boldsymbol{\mu} = b^{'}(\vartheta), \tag{4.2}$$

$$var(\boldsymbol{y}) = \sigma^2 = b^{"}(\vartheta)a(\phi). \tag{4.3}$$

In the Generalized linear model (GLM) frame work, link function is defined to model the relationship between the linear predictor $\boldsymbol{\eta}$ and the expected value $\boldsymbol{\mu}$ of the dependent variable $\boldsymbol{y}$,

$$\boldsymbol{\eta} = g(E[\boldsymbol{y} \mid \boldsymbol{x}]) = \boldsymbol{x}^{'}\boldsymbol{\beta}, \tag{4.4}$$

where $g(\cdot)$ is the link function. The link functions may be arbitrary and different link functions can be applied. A special link function is the canonical link for the exponential family which is given by

$$\boldsymbol{\eta} = \boldsymbol{x}^{'}\boldsymbol{\beta} = \boldsymbol{\vartheta}. \tag{4.5}$$

The finite mixture approach is established on the GLM framework. Grün and Leisch (2008) gave the finite mixture density with K components as

$$f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_k), \tag{4.6}$$

where $\boldsymbol{\Theta}$ denotes the vector of all parameters for the mixture density $f()$, and f_k is the component specific density function which assumed to be univariate and from the exponential family of distributions. The component specific parameters are given by $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^{'}, \phi_k)$, where ϕ_k is the dispersion parameter. The mean of each component is given by

$$\mu_k(\boldsymbol{x}) = g_k^{-1}(\boldsymbol{x}^{'}\boldsymbol{\beta_k}), \tag{4.7}$$

where $g_k()$ is the component specific link function.

For the component weight π_k, it satisfies

$$\sum_{k=1}^{K} \pi_k = 1. \qquad and \qquad \pi_k > 0, \forall k \tag{4.8}$$

4.2.2 The standard EM algorithm for finite mixture model

The EM algorithm is a common approach for finding the ML parameters in the present of incomplete data. In the case of finite mixture model, the missing data is the latent variable z_{ik} which indicates if observation i comes from component k. This means that z_{ik} equals 1 if individual i is from component k and 0 otherwise. If z_{ik} is known, the MLE could be quite easier to compute and the computation is very straight forward. While in most practice problems, the component label indicators z_{ik} are hidden and the EM algorithm could be used to provide iterative steps to maximize the likelihood function. The EM algorithm consists of two steps: E-step and M-step. In the E-step, the expectation of the complete log-likelihood function is computed conditioning on the unobserved data, given the observed data and provisional estimates of parameters. In the M-step, the expected log-likelihood found on the E-step is maximized with respect to the parameters. The E-step and M-step will be iterated until algorithm convergence. The estimation problem is formulated as below.

Let $\{(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)\}$ be the observed data set with y_i the dependent variable and $\boldsymbol{x}_i$ the covariate vector. y_is and $\boldsymbol{x}_i$s are denoted as $\boldsymbol{Y}$ and $\boldsymbol{X}$ respectively. Assume the data is generated from a mixture of K components distribution of GLMs in proportions of $\pi_1, \ldots, \pi_K$. The log likelihood for the parameters $\boldsymbol{\Theta}$ that can be

formed from these data under fixture model (Equation 4.6) is given by

$$\log L(\boldsymbol{\Theta}) = \sum_{i=1}^{n} log(\sum_{k=1}^{K} \pi_k f_k(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k)). \tag{4.9}$$

The unobservable indicator variable z_{ik} is assumed to be i.i.d. multinomial:

$$f(\boldsymbol{z}_i \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{ik}}, \tag{4.10}$$

where the vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik})^T$ and we denote $\boldsymbol{z}$ the matrix of $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$. Further, it is assumed that the y_i given $\boldsymbol{z}_i$ are conditional independent, i.e.,

$$f(y_i \mid \boldsymbol{z}_i) = \prod_{k=1}^{K} f_k(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k)^{z_{ik}}. \tag{4.11}$$

With z_{ik} considered as missing data, the complete-data log likelihood can be formed as

$$\log L_c(\boldsymbol{\Theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \log \pi_k + \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \log f_k(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k). \tag{4.12}$$

It has been shown in Dempster et al. (1977) that the EM algorithm monotonically increases the log likelihood of the observed data. By Jensen's inequality, the log likelihood is proved to be bounded above. The EM algorithm is detailed below.

E-step:

In the E-step, the expected value of Equation 4.12 is taken with respect to $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Theta}^{(l)})$, where $\boldsymbol{\Theta}^{(l)}$ is the current estimate of the parameters. The expectation is called Q function. That is, in the E-step of the lth iteration, we compute

$$Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(l)}) = E(\log L_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Theta}^{(l)})). \tag{4.13}$$

To obtain this expectation, the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{z}$ need to be calculated first

$$f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\Theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} f_k(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k)^{z_{ik}}. \tag{4.14}$$

Using Bayes' rule, the conditional distribution of z_{ik} is derived from Equations 4.11 and 4.14,

$$\hat{p}_{ik} = E(z_{ik} \mid y_i, \boldsymbol{\Theta}) = \frac{\pi_k^{(l)} f_k(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k^{(l)})}{\sum_{j=1}^{K} \pi_j^{(l)} f_j(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_j^{(l)})}. \tag{4.15}$$

M-step:

The M-step on the $(l+1)$th iteration requires to maximize $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(l)})$ with respect to $\boldsymbol{\Theta}$ to get the updated estimation of $\boldsymbol{\Theta}^{(l+1)}$. Replace the non-observed data $\boldsymbol{Z}$ in Equation 4.12 by their current expectation $\hat{p}_{ik}$,

$$E(\log L_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Theta}^{(l)})) = \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{p}_{ik}^{(l)} \log \pi_k + \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{p}_{ik}^{(l)} \log f_k(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k^{(l)}). \tag{4.16}$$

As the cross-derivatives of the two terms on the right side of Equation 4.16 are equal to 0, they can be maximized separately. Let

$$Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(l)}) = Q_1(\pi \mid \boldsymbol{\Theta}^{(l)}) + Q_2(\boldsymbol{\theta} \mid \boldsymbol{\Theta}^{(l)}), \tag{4.17}$$

where

$$Q_1(\boldsymbol{\pi} \mid \boldsymbol{\Theta}^{(l)}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{p}_{ik}^{(l)} \log \pi_k, \tag{4.18}$$

and

$$Q_2(\boldsymbol{\theta} \mid \boldsymbol{\Theta}^{(l)}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{p}_{ik}^{(l)} \log f_k(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k^{(l)}). \tag{4.19}$$

The maximization of Q_1 with respect to $\boldsymbol{\pi}$, under the restriction for the component weights given in Equation 4.8, is obtained by maximizing the Lagrangian function

$$\mathcal{L}(\boldsymbol{\pi}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{p}_{ik}^{(l)} \log \pi_k + \lambda(\sum_{k=1}^{K} \pi_k - 1), \tag{4.20}$$

where λ is the Lagrange multiplier. Taking the derivative of Equation 4.20, we find

$$\frac{\partial \mathcal{L}(\boldsymbol{\pi})}{\partial \pi_k} = \sum_{i=1}^{n} \frac{\hat{p}_{ik}^{(l)}}{\pi_k} + \lambda. \tag{4.21}$$

Setting this to be zero and solving it, we get the update for parameters $(\pi_k^{(l+1)})_{k=1,\cdots,K}$,

$$\pi_k^{(l+1)} = \frac{1}{n} \sum_{t=1}^{n} \hat{p}_{tk}^{(l)}. \qquad \forall k = 1, \ldots, K \tag{4.22}$$

The maximization of Q_2 gives new estimates $\boldsymbol{\theta}^{(l+1)}$. Taking the derivative of Q_2 with respect to $\boldsymbol{\beta}$ (recall $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^{'}, \phi_k)$) and set it to be zero,

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \hat{p}_{ik}^{(l)} \frac{\partial \log f_k(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k^{(l)})}{\partial \boldsymbol{\beta}} = 0. \tag{4.23}$$

Following from the work on the ML fitting of a single GLM (McCullagh and Nelder 1989), Equation 4.23 can be written as

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \hat{p}_{ik}^{(l)} w(\mu_{ik})(y_i - \mu_{ik}) \frac{\partial \eta_i(\mu_{ik})}{\partial \mu_{ik}} \frac{\partial \eta_i(\mu_{ik})}{\partial \boldsymbol{\beta}} = 0, \tag{4.24}$$

where $w(\boldsymbol{\mu})$ is the weight function defined by

$$w(\mu_{ik}) = \frac{1}{(\frac{\partial \eta_i}{\partial \mu_{ik}})^2 V_{ik}}, \tag{4.25}$$

where V_{ik} is the variance function (see Section 4.3.1 for detail) and for the kth component, μ_{ik} is the mean of Y_i. If $(\beta_1, \ldots, \beta_k)$ are independent from each other, then

$$\frac{\partial \eta_i(\mu_{ik})}{\partial \beta_j} = \boldsymbol{x}_i, \qquad if\ j = i \tag{4.26}$$

$$= 0. \qquad otherwise \tag{4.27}$$

Equation 4.24 is reduced to solve

$$\sum_{i=1}^{n} \hat{p}_{ik}^{(l)} w(\mu_{ik})(y_i - \mu_{ik}) \frac{\partial \eta_i(\mu_{ik})}{\partial \mu_{ik}} \boldsymbol{x}_i = 0 \tag{4.28}$$

separately for each β_k to generate the update $\beta_k^{(l+1)}$ $(k = 1, \ldots, K)$.

It can be easily seen that Equation 4.28 has the same form as a single GLM fitted across all observations with prior weights $\hat{p}_{ik}^{(l)}$. Therefore, for each component k, iteratively reweighted lease-squares method proposed by (Nelder and Baker 1972) for GLM can be applied to produce the maximization of L_i with fixed weight $\hat{p}_{ik}^{(l)}$. Let $\boldsymbol{\beta^{(l)}}$ be the current estimation of regression parameters, form the adjusted dependent variable for each component k

$$s_{ik}^{(l)} = \eta_{ik}^{(l)} + (y_i^{(l)} - \mu_{ik}^{(l)}) \frac{\partial \eta_i(\mu_{ik})}{\partial \mu_{ik}}|_{\mu_{ik}=\mu_{ik}^{(l)}}, \tag{4.29}$$

where $\eta_{ik}^{(l)}$ is the current estimate of the linear predictor in Equation 4.4 and $\mu_{ik}^{(l)}$ is the corresponding fitted value derived from the link function. Regressing the adjusted variable $s_{ik}^{(l)}$ on the covariate $\boldsymbol{x}_i$ with weight $W_{ik}^{(l)}$ for each component k, where

$$W(\mu_{ik})_{(l)} = \frac{\hat{p}_{ik}^{(l)}}{(\frac{\partial \eta_i(\mu_{ik})}{\partial \mu_{ik}}|_{\mu_{ik}=\mu_{ik}^{(l)}})^2 V_{ik}^{(l)}}, \tag{4.30}$$

we will get the new estimates of parameters $\boldsymbol{\beta}^{(l+1)}$. This procedure is repeated until changes are sufficiently small. The estimation of dispersion parameter ϕ_k ($k = 1, \ldots, K$) will be introduced in Section 4.3.1.

BIC can be used to determine the optimum number of components.

4.3 Quasi-likelihood approach

Likelihood function is widely used in estimation. It requires strong assumption about the structure of the data. In many cases there is insufficient knowledge regarding the probability model of the data. However we may able to determine some of the characteristics of the data, such as, how the mean is affected by external stimuli and how the variance changes with the mean. In this situation, we may drop the distribution assumption and only model the first two moments. The term quasi-likelihood was first introduced by Robert Wedderburn in 1974 Wedderburn (1974). McCullagh and Nelder (1989) gave the detailed description of the quasi-likelihood function and made it popular.

4.3.1 Independent data

Suppose that $\boldsymbol{y}$ is the vector of response of dimension $n \times 1$, which are independent with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma^2 \boldsymbol{V}(\boldsymbol{\mu})$, where σ^2 may be unknown and $\boldsymbol{V}(\boldsymbol{\mu})$ is made up of known functions. We assume that $\boldsymbol{\mu}$ is a function of covariate, $\boldsymbol{x}$, which is of dimension $n \times p$, and some regression parameters, $\boldsymbol{\beta}$, i.e, $g(\boldsymbol{\mu}) = \boldsymbol{x}\boldsymbol{\beta}'$. Since the components of $\boldsymbol{y}$ are assumed to be independent, $\boldsymbol{V}(\boldsymbol{\mu})$ must be diagonal and can be written as a $n \times n$ matrix

$$\boldsymbol{V}(\boldsymbol{\mu}) = diag(V_1(\mu), \ldots, V_n(\mu)). \tag{4.31}$$

To construct the quasi-likelihood, we start with a single component y of $\boldsymbol{y}$. Under the conditions listed above, the function

$$U = u(\mu, y) = \frac{y - \mu}{\sigma^2 V(\mu)} \tag{4.32}$$

has the following properties same as the traditional score functions

$$\begin{aligned} E(U) &= 0 \\ var(U) &= 1/[\sigma^2 V(\mu)] \\ -E(\frac{\partial U}{\partial \mu}) &= 1/[\sigma^2 V(\mu)]. \end{aligned} \tag{4.33}$$

Expending the first order asymptotic theory embedded in likelihood field, QL could be defined as

$$Q(\mu; y) = \int_y^{\mu} \frac{y - t}{\sigma^2 V(t)} dt, \tag{4.34}$$

which behaves like a log-likelihood function. Since the components of $\boldsymbol{y}$ are independent by assumption, the quasi-likelihood for the complete data is the sum of the individual ones

$$\boldsymbol{Q}(\boldsymbol{\mu}; \boldsymbol{y}) = \sum Q(\mu_i; y_i). \tag{4.35}$$

The quasi-score function in the vector form can be written as

$$\boldsymbol{U}(\boldsymbol{\mu}; \boldsymbol{y}) = \boldsymbol{V}^{-1}(\boldsymbol{\mu}) \frac{\boldsymbol{y} - \boldsymbol{\mu}}{\sigma^2}. \tag{4.36}$$

We will write $\boldsymbol{V}$ instead of $\boldsymbol{V}(\boldsymbol{\mu})$ for simplicity in the following formulas. The ultimate parameter of interest is the regression coefficient $\boldsymbol{\beta}$. By chain rule, we can get the quasi-score function

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{y}) &= \boldsymbol{U}(\boldsymbol{\mu}; \boldsymbol{y}) \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \\ &= \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \boldsymbol{V}^{-1} \frac{\boldsymbol{y} - \boldsymbol{\mu}}{\sigma^2} \\ &= \boldsymbol{D}^T V^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) / \sigma^2, \end{aligned} \tag{4.37}$$

where $\boldsymbol{D}$ has order of $n \times p$, with component $D_{ir} = \partial \mu_i / \partial \beta_r$, the derivatives of $\boldsymbol{\mu}(\boldsymbol{\beta})$ with respect to the parameters. The quasi-likelihood estimating equations for the regression parameters $\boldsymbol{\beta}$ can be given by setting: $U(\hat{\boldsymbol{\beta}}) = 0$.

By adopting the Fisher Scoring algorithm, we could solve for $\hat{\boldsymbol{\beta}}$. The covariance matrix of $\boldsymbol{U}(\boldsymbol{\beta})$ is also the negative expected value of $\partial \boldsymbol{U}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$

$$\begin{aligned} \boldsymbol{i}_\beta &= cov(\boldsymbol{U}(\boldsymbol{\beta};\boldsymbol{y})) \\ &= cov(\boldsymbol{D}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/\sigma^2) \\ &= \boldsymbol{D}^T\boldsymbol{V}^{-1}cov(\boldsymbol{y})\boldsymbol{V}^{-1}\boldsymbol{D}/\sigma^4 \\ &= \boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}/\sigma^2. \end{aligned} \tag{4.38}$$

This matrix serves the same purposes like the Fisher information for likelihood function. In particular, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$cov(\hat{\boldsymbol{\beta}}) \simeq \boldsymbol{i}_\beta^{-1} = \sigma^2(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}. \tag{4.39}$$

The iterative procedures to solve for $\hat{\boldsymbol{\beta}}$ generated by the Newton-Raphson method are given below: beginning with and arbitrary value $\hat{\boldsymbol{\beta}}_0$ sufficiently close to $\hat{\boldsymbol{\beta}}$, the sequence of parameter estimates is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 &= \hat{\boldsymbol{\beta}}_0 + \boldsymbol{i}_\beta^{-1}\boldsymbol{U}(\hat{\boldsymbol{\beta}}_0) \\ &= \hat{\boldsymbol{\beta}}_0 + (\hat{\boldsymbol{D}}_0^T\hat{\boldsymbol{V}}_0^{-1}\hat{\boldsymbol{D}}_0)^{-1}\hat{\boldsymbol{D}}_0^T\hat{\boldsymbol{V}}_0^{-1}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0). \end{aligned} \tag{4.40}$$

Iterating the above the procedure until convergence occurs, the quasi-likelihood estimate $\hat{\boldsymbol{\beta}}$ may be obtained.

As for the estimation of dispersion parameter σ^2, there is no equivalent to an Maximum Likelihood estimate. Therefore the moment estimator based on the resid-

ual vector $(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$, is used. The generalized Pearson-Chi-square statistic takes the form

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{var(y_i)} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2 V_i(\hat{\mu}_i)} \overset{D}{\sim} \chi^2_{n-p}. \tag{4.41}$$

where χ^2_{n-p} is the generalized Pearson statistic with $n-p$ degree of freedom. Taking expectation of both sides of equation 4.41,

$$E(\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2 V_i(\hat{\mu}_i)}) = n - p, \tag{4.42}$$

we get

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{\chi^2_{n-p}}{n-p}. \tag{4.43}$$

4.3.2 Dependent data

There are many situations where the dependence relationships among the data are too significant to ignore. Longitudinal data and time series are examples of dependent data, in which repeated measures made on the same subjects over time are usually positively related. While the data are correlated, the most important change is in the matrix $\boldsymbol{V}(\boldsymbol{\mu})$, where $\boldsymbol{V}(\boldsymbol{\mu})$ is a symmetric positive-definite $n \times n$ matrix of known functions $\boldsymbol{V}_{ij}(\boldsymbol{\mu})$, no longer diagonal. Liang and Zeger (1986) pioneered at applying the quasi-likelihood approach to longitudinal data analysis and proposed generalized estimated equations (GEE) approach.

Suppose that there is a longitudinal data set $\{y_{it}, \boldsymbol{x}_{it}\}$ with mean $\mu_{it} = E(y_{it})$ and link function $g(\mu_{it}) = \boldsymbol{x}_{it}^T\boldsymbol{\beta}$ for i-th subject measured at time point t, $t = 1, 2, \ldots, n_i$ and subjects $i = 1, 2, \ldots, n$. Here y_{it} is the dependent variable and $\boldsymbol{x}_{it}$ is the independent variable of dimension $p \times 1$ at time point t. Let $\boldsymbol{y}_i$ be a $n_i \times 1$ vector $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$, with mean $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \times, \mu_{in_i})^T$ and covariance matrix $var(\boldsymbol{y}_i) = \phi \boldsymbol{V}(\boldsymbol{\mu_i})$ where $\boldsymbol{y}_i$ are independent of each other for $i = 1, 2, \ldots, n$. In Liang and Zeger (1986)'s GEE approach, a working correlation matrix $\boldsymbol{R}(\boldsymbol{\alpha})$ is applied. $\boldsymbol{R}(\boldsymbol{\alpha})$ is assumed to be a symmetric matrix which fulfills the requirement of being a correlation matrix and can be fully characterized by a vector parameter $\boldsymbol{\alpha}$. Combining all these assumptions,

$$var(\boldsymbol{y}_i) = \boldsymbol{\Sigma}_i = \phi \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{A}_i^{\frac{1}{2}}, \tag{4.44}$$

where $\boldsymbol{A}_i = diag(V(\mu_{i1}), V(\mu_{i2}), \ldots, V(\mu_{in_i}))$. The GEE is defined to be

$$\begin{aligned} \boldsymbol{U}(\beta) &= \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ &= 0, \end{aligned} \tag{4.45}$$

where the matrix $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i \backslash \partial \boldsymbol{\beta}$. The model based covariance is given by the inverse

of I_0,

$$\begin{aligned} I_0 &= E(-\frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}) \\ &= -\sum_{i=1}^{n} E(\frac{\partial(\boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1})}{\boldsymbol{\beta}}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) + \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial(\boldsymbol{y}_i - \boldsymbol{\mu}_i)}{\partial \boldsymbol{\beta}}) \\ &= \sum_{i=1}^{n} E(\boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}) \\ &= \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i \end{aligned} \tag{4.46}$$

Given current estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$, the Newton Raphson algorithm to estimate $\hat{\boldsymbol{\beta}}$ at time t is

$$\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^{t} + \boldsymbol{I}_0^{-1}(\hat{\boldsymbol{\beta}}^{t}) \boldsymbol{U}(\hat{\boldsymbol{\beta}}^{t}). \tag{4.47}$$

At a given iteration the correlation parameters $\boldsymbol{\alpha}$ and scale parameter ϕ can be estimated from the current Pearson residuals defined by

$$\hat{r}_{it} = \frac{(y_{it} - \mu_{it})}{\sqrt{V(\mu_{it})}}. \tag{4.48}$$

The estimation of scale parameter is

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{n_i} \hat{r}_{it}^2}{N - p}, \tag{4.49}$$

where $N = \sum n_i$, and the simple function to estimate $\boldsymbol{\alpha}$ is

$$\hat{R_{uv}} = \sum_{i=1}^{n} \hat{r}_{iu} \hat{r}_{iv} / (N - p). \tag{4.50}$$

Iterating this algorithm until convergence will get the estimate of regression parameters $\hat{\boldsymbol{\beta}}$.

There are two classical ways of estimating the covariance $cov(\hat{\boldsymbol{\beta}})$. One is model based estimation: $cov(\hat{\boldsymbol{\beta}})_m = \boldsymbol{I}_0^{-1}$, which consistently estimates $cov(\hat{\boldsymbol{\beta}})$ while the mean model and the working correlation are correct. The other is empirical estimate obtained by sandwich estimate. The sandwich estimator was first proposed by Huber (1967)and White (1980); Liang and Zeger (1986) applied it to longitudinal data,

$$cov(\hat{\boldsymbol{\beta}})_e = \boldsymbol{I}_0^{-1}\boldsymbol{I}_1\boldsymbol{I}_0^{-1}, \tag{4.51}$$

where

$$\begin{aligned} \boldsymbol{I}_1 &= cov(\boldsymbol{U}(\hat{\boldsymbol{\beta}})) \\ &= cov(\sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)) \\ &= \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1} cov(\boldsymbol{y}_i) \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1}. \end{aligned} \tag{4.52}$$

Here $cov(\hat{\boldsymbol{\beta}})_e$is a consistent estimate of $cov(\hat{\boldsymbol{\beta}})$ even if the working correlation is misspecified, i.e. if $cov(\boldsymbol{y}_i) \neq \boldsymbol{\Sigma}_i$. In practice, $cov(\boldsymbol{y}_i)$ is replaced by $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'$, i.e.,

$$\boldsymbol{I}_1 = \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1}. \tag{4.53}$$

4.3.3 Morkov regression models for time series

Zeger and Qaqish (1988) studied Markov regression models for time series using quasi-likelihood approach. They defined y_t to be an outcome time series and $\boldsymbol{x}_t$ an $m \times 1$ vector of covariates for $t = -p+1, \ldots, 0, 1, \ldots, n$. Let $\boldsymbol{B}_t$ be the present and past covariates and past observations at time t, i.e.,

$$\boldsymbol{B}_t = \{\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{-p+1}, y_{t-1}, y_{t-2}, \ldots, y_{-p+1}\}. \tag{4.54}$$

Define

$$\mu_t = E(y_t \mid \boldsymbol{B}_t) \qquad and \qquad \nu_t = var(y_t \mid \boldsymbol{B}_t). \tag{4.55}$$

They assume

$$g(\mu_t) = \eta = \boldsymbol{x}_t^{'}\boldsymbol{\beta} + \sum_{i=1}^{q} \lambda_i h_i(\boldsymbol{B}_t), \tag{4.56}$$

where g is a "link" function, η is a linear predictor, and h_i's are functions of the past outcomes and the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^{'}$ are to be estimated. As in quasi-likelihood approach, they further assume that

$$\nu_t = var(y_t \mid \boldsymbol{B}_t) = V(\mu_t) \cdot \phi, \tag{4.57}$$

where V is a variance function and ϕ is an unknown scale parameter. Comparing to the formulation defined by (McCullagh and Nelder 1989) for the independent data,

this formulation has the same format except that conditional rather than marginal moments are modelled.

Let $\boldsymbol{\gamma}^{'} = (\boldsymbol{\beta}^{'}, \boldsymbol{\lambda}')$. These parameters are to be estimated by a quasi-likelihood approach so that $\boldsymbol{\gamma}$ is the root of the log-quasi-likelihood estimating equation

$$\begin{aligned} U(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} \frac{\partial \mu_t}{\partial \boldsymbol{\gamma}} \nu_t^{-1}(y_t - \mu_t) \\ &= \sum_{i=1}^{n} \frac{\partial \mu_t}{\partial \eta} \frac{\partial \eta}{\partial \boldsymbol{\gamma}} \nu_t^{-1}(y_t - \mu_t) \\ &= \sum_{i=1}^{n} \frac{\partial \mu_t}{\partial \eta} \nu_t^{-1} \boldsymbol{M}_t(y_t - \mu_t), \end{aligned} \tag{4.58}$$

where $\boldsymbol{M}_t^{'} = (\boldsymbol{x}_t^{'}, h_1(\boldsymbol{B}_t), \ldots, h_q(\boldsymbol{B}_t))$. While using canonical link, $\frac{\partial \mu_t}{\partial \eta} = \frac{\partial \mu_t}{\partial \theta}$ can be derived from Equation 4.5. Combining with Equations 4.2 and 4.3, we get

$$\frac{\partial \mu_t}{\partial \eta} = \frac{\partial \mu_t}{\partial \theta} = \frac{\partial (b_t(\theta))}{\partial \theta} = b_t^{''}(\theta) = \frac{\nu_t}{\phi}. \tag{4.59}$$

Therefore, with canonical link, equation 4.58 reduces to

$$U(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \frac{\nu_t}{\phi} \nu_t^{-1} \boldsymbol{M}_t(y_t - \mu_t) = 0 \tag{4.60}$$

$$= \sum_{i=1}^{n} \boldsymbol{M}_t(y_t - \mu_t) = 0. \tag{4.61}$$

Equation 4.60 could be solved by iteratively reweighted least squares algorithm.

4.3.4 Asymptotic theory

The asymptotic theorem on maximum likelihood estimation for general stochastic processes with discrete time was given by Kaufmann (1987). They defined $\{y_t, t = 1, 2, \ldots\}$ be such a process on a probability space $(\Omega, \mathfrak{F}, P)$. Let $\mathfrak{F}_t$ denote the σ-field generated by the first t observations $y_1, \ldots, y_t$, and $\mathfrak{F}_0 = \{\emptyset, \Omega\}$. Assume the probability measure P belong to a parametric family $P_\beta, \beta \in B$, where the parameter space B is an open subset of $\mathbb{R}^p, p \in \mathbb{N}$. For fixed t, let the projections $\{P_{t,\beta}, \beta \in B\}$ on the first t observations be mutually absolutely continuous. Assume the likelihood to be two times continuously differentiable. Denote $l_t(\boldsymbol{\beta}), s_t(\boldsymbol{\beta}), -\boldsymbol{H}_t(\boldsymbol{\beta})$ to be the log-likelihood and its first and second derivatives respectively, and define $a_t(\boldsymbol{\beta}) = s_t(\boldsymbol{\beta}) - s_{t-1}(\boldsymbol{\beta})$. The following assumption is made for the asymptotic theorem.

Assumption A.

1. The score function $\{s_t\}$, evaluated at $\boldsymbol{\beta}$, is a square integrable zero mean martingale with respect to $\{\mathfrak{F}\}$.

2. With some nonrandom nonsingular norming sequence $\{\boldsymbol{A}_t^{1/2}\}$, the conditional information $\boldsymbol{G}_t(\boldsymbol{\beta}) = \sum_1^t cov_{\boldsymbol{\beta}}(a_s(\boldsymbol{\beta}) \mid \mathfrak{F}_{t-1})$ converges to a random a.s. positive definite matrix,

$$\boldsymbol{A}_t^{1/2} \boldsymbol{G}_t(\boldsymbol{\beta}) \boldsymbol{A}_t^{-T/2} \xrightarrow{p} \boldsymbol{V}(\boldsymbol{\beta}). \tag{4.62}$$

3. The Lindeberg condition holds, i.e., for any $\varepsilon > 0$,

$$\sum_{1}^{t} E(a_s' \boldsymbol{A}_t^{-1} a_s I_{ts}(\varepsilon) \mid \mathfrak{F}_{s-1}) \xrightarrow{p} 0, \tag{4.63}$$

where $I_{ts}(\varepsilon)$ is the indicator of $\{a_s' \boldsymbol{A}_t^{-1} a_s \geq \varepsilon^2\}$.

4. The continuity condition

$$\sup_{\tilde{\boldsymbol{\beta}} \in N_t(\delta)} \|\boldsymbol{A}_t^{-1/2}(\boldsymbol{H}_t(\tilde{\boldsymbol{\beta}}) - \boldsymbol{G}_t)\boldsymbol{A}_t^{-T/2}\| \xrightarrow{p} 0, \tag{4.64}$$

with $\boldsymbol{N}_t(\delta) = \{\tilde{\boldsymbol{\beta}} : \|\boldsymbol{A}_t^{T/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\| \leq \delta\}$, holds for any $\delta > 0$.

Theorem 1. Under assumption N, the probability that a locally unique MLE exists converges to one. Moreover, there exists a sequence $\{\hat{\boldsymbol{\beta}}_t\}$ of MLE's which is consistent and asymptotically normal

$$\boldsymbol{G}_t^{T/2}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}) \rightarrow N(0, I), \tag{4.65}$$

with an appropriate square root $\boldsymbol{G}_t^{T/2}$.

The proof of the above theorem can be found in reference (Kaufmann 1987).

4.4 A mixture of GLM with Markov process for cluster time series

In this study, finite mixture GLM and Markov regression models were combined to develop model-based clustering algorithm to cluster the time series.

4.4.1 Model specification

We propose the following mixture Markov model to analysis mixture time series. Let $\boldsymbol{Y} = (\boldsymbol{y}_1', \boldsymbol{y}_2', \ldots, \boldsymbol{y}_n')'$ be a set of n independent time series. Assume each time series is observed at time points $t = -q+1, \ldots, 0, 1, \ldots, n_i$, on a probability space $(\Omega, \mathfrak{F}, P)$ and the time series is denoted as $\boldsymbol{y}_i = (y_{(i,-q+1)}, \ldots, y_{(i,n_i)})'$. Furthermore, the covariates for the ith time series are measured at each time point denoted as $\boldsymbol{x}_{it}$. For the ith time series, let the collection of present and past covariates and past observations for time t be denoted as $\boldsymbol{B}_{it} = (\boldsymbol{x}'_{(i,t)}, \boldsymbol{x}'_{(i,t-1)}, \ldots, \boldsymbol{x}'_{(i,-q+1)}, y_{(i,t-1)}, y_{(i,t-2)}, \ldots, y_{(i,-q+1)})'$. Let $\mathfrak{F}_t$ denote the σ-field generated by the first t observations $y_{\cdot 1}, \ldots, y_{\cdot t}$ for each time series, and $\mathfrak{F}_0 = \{\emptyset, \Omega\}$. Assume the probability measure P belong to a parametric family $P_\beta, \beta \in S$, where the parameter space S is an open subset of $\mathbb{R}^p, p \in \mathbb{N}$. For fixed t, let the projections $\{P_{t,\beta}, \beta \in S\}$ on the first t observations be mutually absolutely continuous. Assume the likelihood to be two times continuously differentiable.

Assume the time series has K mixture components. For the kth component, conditional on the present and past covariates and past observations, the observation $y_{(i,t)}$, abbreviated as y_{it}, is assumed to follow an exponential family distribution denoted as $f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k)$. Each time series is randomly drawn from one of the K components with probability π_k and $\sum_{k=1}^{K} \pi_k = 1$. The quasi-likelihood formulated as the product of all conditional probabilities for a complete time series from the kth

mixture component is defined as:

$$\tilde{P}^{(k)}(\boldsymbol{y}_i \mid \boldsymbol{B}_i, \boldsymbol{\theta}_k) = \prod_{t=1}^{n_i} f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k), \tag{4.66}$$

where $\boldsymbol{B}_i = \{\boldsymbol{B}_{it}, t = 1, \ldots, n_i\}$. We further define $\mu_{it}^{(k)} = E(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k)$ and $\nu_{it}^{(k)} = \text{var}(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k)$. It is assumed that

$$g(\mu_{it}^{(k)}) = \eta_{it}^{(k)} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta}_k + \sum_{j=1}^{q} \lambda_{kj} h_{kj}(\boldsymbol{B}_{it}), \tag{4.67}$$

where g is a "link" function, h_{kj}'s are functions of the past observations, and $\boldsymbol{\theta}_k$ encompasses all the regression coefficients $(\boldsymbol{\beta}_k, \lambda_{k1}, \ldots, \lambda_{kq})'$. It is further assumed that $\nu_{it}^{(k)} = V(\mu_{it}^{(k)}) \cdot \phi$, where V is a variance function and ϕ is an unknown scale parameter. As the component membership is unknown, the overall quasi-likelihood is formulated as a mixture of K different quasi-likelihoods:

$$\tilde{P}(\boldsymbol{y}_i \mid \boldsymbol{B}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \tilde{P}^{(k)}(\boldsymbol{y}_i \mid \boldsymbol{B}_i, \boldsymbol{\theta}_k), \tag{4.68}$$

where $\boldsymbol{\Theta}$ denotes the vector of all parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\} \cup \{\pi_1, \ldots, \pi_K\}$. This setup is an extension of the quasi-likelihood formulation of a single Markov regression model (Zeger and Qaqish 1988) to a mixture of Markov regression models.

The overall quasi-likelihood of n independent time series can be formulated as:

$$\tilde{P}(\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Theta}) = \prod_{i=1}^{n} \{\sum_{k=1}^{K} \pi_k \prod_{t=1}^{n_i} f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k)\}, \tag{4.69}$$

where $\boldsymbol{B} = \{\boldsymbol{B}_{it}, i = 1, \ldots, n, t = 1, \ldots, n_i\}$. Then the log quasi-likelihood given the data set and covariates is given by

$$\log \tilde{L}(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{B}) = \sum_{i=1}^{n} \log\{\sum_{k=1}^{K} \pi_k \prod_{t=1}^{n_i} f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k)\}. \tag{4.70}$$

Each time series $\boldsymbol{y}_i$ follows a Markov model of order q with a semi-positive definite covariance matrix. It is important to note that the time series are in general non-stationary because of the exogenous variables. The mean and variance at each time point are influenced by the time varying exogenous variables. As a special example, if y_{it} follows a normal distribution, and the link function is the identity function, $h_{kj}(\boldsymbol{B}_{it}) = y_{(i,t-j)} - \boldsymbol{x}'_{(i,t-j)}\boldsymbol{\beta}_k$, in Equation 4.67, then the proposed model is an autoregressive model of order q. The partial autocorrelation will cut off after lag q and the autocorrelation will follow a geometric decay given the parameters satisfy the stationary condition for the underlying autoregressive model.

4.4.2 EM algorithm for the GLM with Markov process

Similarly to the EM algorithm for the finite mixture model, the EM algorithm applied to our GLM with Markov process can be detailed below.

Suppose the memberships are known, let $\boldsymbol{Z}$ be the matrix of membership indicators with $\boldsymbol{Z} = \{z_{ik}\}, i = 1, \ldots, n,\ k = 1, \ldots, K$. If the ith time series belongs to the kth component, then $z_{ik} = 1$ and $z_{ij} = 0, \forall j \neq k$. With the complete membership

data, the complete quasi-likelihood can be defined as:

$$\tilde{P}_c(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{\Theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\prod_{t=1}^{n_i} f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k)]^{z_{ik}}. \tag{4.71}$$

Based on the assumptions made in Section 4.4.1, the complete log quasi-likelihood follows directly from Equation 4.71

$$\log \tilde{L}_c(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{B}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{t=1}^{n_i} z_{ik} \log f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k). \tag{4.72}$$

The proposed EM algorithm consists of the E-step and M-step with both steps modified to deal with the formulation of quasi-likelihood instead of true likelihood.

E-step:

In the E-step of the lth iteration, we compute the Q function which is defined as,

$$Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(l)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{t=1}^{n_i} E[z_{ik} \log f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k) | y_{it}, \boldsymbol{B}_{it}, \boldsymbol{\theta}_k^{(l)}]. \tag{4.73}$$

As shown in the expression of Q, we compute the expected complete conditional loglikelihood locally at each time point, conditioning on the observed y_{it} and $\boldsymbol{B}_{it}$. This ensures the accent property of the algorithm. By Bayes' rule, the conditional local expectation of the membership indicator z_{ik} is given by

$$\hat{p}_{ikt}^{(l)} = E(z_{ik} \mid \boldsymbol{y}_{it}, \boldsymbol{B}_{it}, \boldsymbol{\theta}_k^{(l)}) = \frac{\pi_k^{(l)} f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k^{(l)})}{\sum_{j=1}^{K} \pi_j^{(l)} f_j(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_j^{(l)})}. \tag{4.74}$$

Plug $\hat{p}_{ikt}^{(l)}$ into Equation 4.73, we have $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(l)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{t=1}^{n_i} \hat{p}_{ikt}^{(l)} \log f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k^{(l)})$.

M-step:

Replace the non-observed data $\boldsymbol{Z}$ in Equation 4.72 by their current expectation $\hat{p}_{ik}$,

$$\begin{aligned} E(\log L_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{\Theta}^{(l)})) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{p}_{ik}^{(l)} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{n_i} \hat{p}_{ik}^{(l)} f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k^{(l)}) \\ &= Q_1(\pi \mid \boldsymbol{\Theta}^{(l)}) + Q_2(\boldsymbol{\theta} \mid \boldsymbol{\Theta}^{(l)}), \end{aligned} \tag{4.75}$$

where

$$Q_1(\pi \mid \boldsymbol{\Theta}^{(l)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{p}_{ik}^{(l)} \log \pi_k, \tag{4.76}$$

and

$$Q_2(\pi \mid \boldsymbol{\Theta}^{(l)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{n_i} \hat{p}_{ik}^{(l)} f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k^{(l)}). \tag{4.77}$$

Maximizing Q_1 with respect to $\boldsymbol{\pi}$ yields

$$\pi_k^{(l+1)} = \frac{1}{n}\sum_{t=1}^{n} \hat{p}_{tk}^{(l)} \qquad \forall k = 1, \ldots, K \tag{4.78}$$

Let $\hat{p}_{ik}^{(l)} = \frac{1}{n_i}\sum_{t=1}^{n_i} \hat{p}_{ikt}^{(l)}$ denote the pooled estimate of the membership probability across all time points for the ith time series. Taking the derivative of Q_2 with respect to $\boldsymbol{\theta}$ and set it to be zero,

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{n_i} \hat{p}_{ik}^{(l)} \frac{\partial \log f_k(y_{it} \mid \boldsymbol{B}_{it}, \boldsymbol{\theta}_k^{(l)})}{\partial \boldsymbol{\theta}} = 0. \tag{4.79}$$

As different from the maximization procedure of Q_2 in Section 2.2, QL approach is adopted to estimate the parameters $\boldsymbol{\theta}$ instead of the ordinary ML, which yields,

$$\begin{aligned} U(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{p}_{ik}^{(l)}\sum_{t=1}^{n_i}\frac{w(\mu_{it}^{(k)})}{\phi}(y_{it}-\mu_{it}^{(k)})\frac{\partial\eta_i(\mu_{it}^{(k)})}{\partial\mu_{it}^{(k)}}\frac{\partial\eta_i(\mu_{it}^{(k)})}{\partial\boldsymbol{\theta}} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{p}_{ik}^{(l)}\sum_{t=1}^{n_i}\frac{w(\mu_{it}^{(k)})}{\phi}(y_{it}-\mu_{it}^{(k)})\frac{\partial\eta_i(\mu_{it}^{(k)})}{\partial\mu_{it}^{(k)}}\boldsymbol{M}_{it} \\ &= 0. \end{aligned} \tag{4.80}$$

For each mixture component, the estimating score equation takes the form

$$U(\boldsymbol{\theta}_k)=\sum_{i=1}^{n}\hat{p}_{ik}^{(l)}\sum_{t=1}^{n_i}w(\mu_{it}^{(k)})(y_{it}-\mu_{it}^{(k)})\frac{\partial\eta_{it}^{(k)}}{\partial\mu_{it}^{(k)}}\boldsymbol{M}_{it}^{(k)}=0, \tag{4.81}$$

where $\boldsymbol{M}_{it}^{(k)'}=(\boldsymbol{x}_{it}',h_{k1}(\boldsymbol{B}_{it}),\ldots,h_{kq}(\boldsymbol{B}_{it}))$, and $w(.)$ is a weight function defined by

$$w(\mu_{it}^{(k)})=\frac{1}{(\frac{\partial\eta_{it}^{(k)}}{\partial\mu_{it}^{(k)}})^2 v_{it}^{(k)}}. \tag{4.82}$$

If $h_{kj}(\boldsymbol{B}_{it})$ does not depend on $\boldsymbol{\beta}$, iteratively reweighted least square method shown in Section 2.2 can be used to obtain the estimation of $\boldsymbol{\theta}$. Otherwise, a second level of iteration is needed. For example,

$$h_{kj}(\boldsymbol{B}_{it})=g(y_{i,t-j})-\boldsymbol{x}_{i,t-j}\boldsymbol{\beta} \tag{4.83}$$

In this case

$$g(\mu_{it}^{(k)})=\tilde{\boldsymbol{x}}_{it}'\boldsymbol{\beta}_k+\lambda_1 g(y_{i,t-1})+\cdots+\lambda_q g(y_{i,t-q}) \tag{4.84}$$

where $\tilde{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - \lambda_1 \boldsymbol{x}_{i,t-1} - \cdots - \lambda_q \boldsymbol{x}_{i,t-q}$. Let $\tilde{\boldsymbol{X}}_{it} = (\tilde{\boldsymbol{x}}'_{it}, g(y_{i,t-1}), \ldots, g(y_{i,t-q}))'$. The model can be reformulated as regressing y_{it} on $\tilde{\boldsymbol{X}}_{it}$ and the iteration procedure is as follows:

1. At the lth iteration, calculate $\tilde{\boldsymbol{x}}_{it}^{(l)}$.

2. Estimate the parameters $\boldsymbol{\theta}^{(l+1)}$ by iteratively reweighted least square method.

3. Repeat steps (1) and (2) until convergence to obtain the updated values for $\boldsymbol{\theta}^{(l+1)}$.

When the number of components K is unknown, the optimal K can be determined by fitting mixture time series with different numbers of components and choosing the model with BIC.

4.4.3 The asymptotic distribution of the parameters

Rewrite the score function 4.81 for the kth component as

$$U_t^{(k)} = \sum_{i=1}^{n} \hat{p}_{ik} \sum_{t=1}^{n_i} w(\mu_{it}^{(k)})(y_{it} - \mu_{it}^{(k)}) \frac{\partial \eta_{it}^{(k)}}{\partial \mu_{it}^{(k)}} \boldsymbol{M}_{it}^{(k)} \tag{4.85}$$

Since $E[\hat{p}_{ik}(y_{i,t+1} - \mu_{i,t+1}^{(k)}) \mid \mathfrak{F}_t] = 0$, $E[U_{t+1}^{(k)}(\boldsymbol{\beta_k}) \mid \mathfrak{F}_t] = U_t^{(k)}(\boldsymbol{\beta}_k)$. This shows that the process $U_t^{(k)}(\boldsymbol{\beta}_k)$ coupled with the filtration $\mathfrak{F}_t, t = 1, \ldots, N$, forms a martingale with zero mean.

Define the conditional variability matrix of the kth component of the mixture Markov model as

$$\boldsymbol{G}_t^{(k)} = \sum_{i=1}^{n} z_{ik} \sum_{t=1}^{n_i} w(\mu_{it}^{(k)}) \boldsymbol{M}_{it} \boldsymbol{M}_{it}^{'} \tag{4.86}$$

$$\textit{and} \qquad \boldsymbol{G}^{(k)} = \lim_{N \to \infty} \frac{1}{N} \boldsymbol{G}_t^{(k)} \tag{4.87}$$

Let $a_t^{(k)}(\beta) = U_t^{(k)}(\boldsymbol{\beta}) - U_{t-1}^{(k)}(\boldsymbol{\beta})$ and $\boldsymbol{H}_t^{(k)}(\boldsymbol{\beta}_k) = -\nabla U_t^{(k)}(\boldsymbol{\beta}_k)$.

Assumption B.

1. Assume this martingale is square integrable.

2. Assume the limit of the conditional information matrix $\boldsymbol{G}^{(k)}$ exists and it is positive definite

3. Assume with some nonrandom nonsingular norming sequence $\boldsymbol{A}_t^{1/2}$, the Lindeberg condition, i.e., Equation 4.63 holds.

4. Assume the continuity condition, i.e, Equation 4.64 holds.

Remarks: Assumption B.2 is similar to the assumption of non-singular design matrix for classical linear regression. According to Lemma 3 in (Kaufmann 1987), under this assumption, the unconditional information matrix of the joint likelihood of the time series is assured to be positive definite. Therefore, Assumption B.2 is equivalent to Assumption A.2.

Proposition: Let $\boldsymbol{Y} = (\boldsymbol{y}_1', \boldsymbol{y}_2', \ldots, \boldsymbol{y}_n')'$ be a set of n independent time series satisfying all the assumptions described in Section 4.4.1 and Assumption B. Let $\hat{\boldsymbol{\theta}}_k$ be the quasi-likelihood estimates for the regression parameters of the kth component mixture Markov model. Assume the limiting matrix $\boldsymbol{G}^{(k)}$ exists and it is positive definite. Then according to Theorem 1 and under the regularity conditions for the conditional density of y_{it} given $\boldsymbol{B}_{it}$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k) \xrightarrow{D} N(0, \boldsymbol{G}^{(k)^{-1}}). \tag{4.88}$$

Proof. Expanding the score equation 4.81 for the mixture Markov model around the true $\boldsymbol{\theta}^{(k)}$ in a Taylor series, we have

$$\mathbf{0} = U(\hat{\boldsymbol{\theta}}_k) \approx U(\boldsymbol{\theta}_k) + \frac{\partial U(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k) + o_p(1). \tag{4.89}$$

Rearranging the terms leads to

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k) = [-\frac{1}{N}\frac{\partial U(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k}]^{-1}\frac{1}{\sqrt{N}}U(\boldsymbol{\theta}_k) + o_p(1). \tag{4.90}$$

Applying the martingale central limit theorem, we have

$$\frac{1}{\sqrt{N}}U(\boldsymbol{\theta}_k) \xrightarrow{d} N(0, \boldsymbol{G}^{(k)}). \tag{4.91}$$

Applying the law of large numbers to the derivative matrix of the score vector yields

$$-\frac{1}{N}\frac{\partial U(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k} \xrightarrow{P} \boldsymbol{G}^{(k)}. \tag{4.92}$$

By Equations 4.91 and 4.92, and applying Slutsky's theorem, the limiting distribution of the quasi-likelihood estimator follows. □

The limiting covariance matrix can be estimated by

$$\hat{\boldsymbol{G}}^{(k)} = \frac{1}{N}\sum_{i=1}^{n}\hat{p}_{ik}\sum_{t=1}^{n_i}\hat{w}(\mu_{it}^{(k)})\boldsymbol{M}_{it}\boldsymbol{M}_{it}^{'}, \tag{4.93}$$

where $\hat{p}_{ik}$ and $\hat{w}(\mu_{it}^{(k)})$ are the limiting values of the estimates when the algorithm converges. The estimation of the scale parameter ϕ for cluster k is

$$\hat{\phi}_k = \frac{\sum_{i=1}^{n}\hat{p}_{ik}\sum_{t=1}^{n_i}\hat{r}_{ikt}^2}{\sum_{i=1}^{n}(\hat{p}_{ik}n_i) - s}, \tag{4.94}$$

where s is the number of parameters in $\boldsymbol{\theta}_k$, and $\hat{r}_{ikt}$ is the Pearson residuals defined by $\hat{r}_{ikt} = (y_{it} - \mu_{it}^{(k)})/[V(\mu_{it}^{(k)})]^{\frac{1}{2}}$.

4.5 Experimental results for simulated datasets

In order to evaluate the performance of the proposed mixture Markov regression method, we conducted simulations on the following three Markov regression models given by Zeger and Qaqish (1988). We consider times series of model (1) with counts formulated as conditional Poisson distribution, model (2) with counts formulated as conditional gamma distribution and model (3) with binary outcomes formulated as conditional binomial distribution. The conditional regression models of the kth

component time series under model (1) and (2) are:

$$\log(\mu_t^{(k)}) = \boldsymbol{x}_t'\boldsymbol{\beta}^{(k)} + \sum_{j=1}^{q} \lambda_j^{(k)}[\log(y_{t-j}^*) - \boldsymbol{x}_{t-j}'\boldsymbol{\beta}^{(k)}], \tag{4.95}$$

where $y_{t-1}^* = max(y_{t-1}, c)$, $0 < c < 1$, with c set to be 0.05 in the simulation, p is the dimension of $\boldsymbol{\beta}^{(k)}$, and q is the number of λ_js. The predictors $\boldsymbol{x}$ were generated from uniform distribution on the interval $[0, 1]$. For time series in model (3), the conditional regression model is:

$$logit(\mu_t^{(k)}) = \boldsymbol{x}_t'\boldsymbol{\beta}^{(k)} + \sum_{j=1}^{q} \lambda_j^{(k)} y_{t-j}, \tag{4.96}$$

and the predictors $\boldsymbol{x}$ were generated from standard normal distribution $N(0, 3)$. In each simulation, the data set consists of $m = 30$ time series of length n, where $n = 50$ or 100. The time series were simulated according to a mixture of $K = 2$ or $K = 3$ components. For each of the three conditional models, different combinations of n, p, q and K were tested and each scenario was repeated 1000 times. The true parameters used in the simulations were given in Table 4.3 and Table 4.4-4.9 provided the quasi-likelihood estimates and the corresponding standard errors.

It is observed that the proposed quasi-likelihood method provides accurate estimation for the true model parameters in all simulation settings. For all the three distributions, the standard errors decrease as the number of observations n in each time series increases. The magnitude of autoregressive parameters λ didn't significantly affect the sizes of the standard errors under the conditional Poisson and

the conditional Gamma models. However it greatly affected the sizes of the standard errors under the conditional Binomial distribution model. As (λ_1, λ_2) increase from (0.10, 0.15) to $(0.60, 0.65)$, the standard errors have large increase accordingly. This is because in our simulated conditional Binomial model, the autoregressive term λ_j multiplies directly with y_{t-j}, whereas in the simulated conditional Poisson and Gamma model, λ_j multiplies with transformed past outcomes in the form of $\log(y^*_{t-j}) - \boldsymbol{x}'_{t-j}\boldsymbol{\beta}$. The latter form of autoregressive term seems to have less influence on the variability of the estimates that the former form. The standard errors under the conditional Gamma model was also affected by the value of the dispersion parameter ϕ. The standard errors are larger when ϕ is greater than 1.

Figure 4.1 gave an example of a simulated dataset for the Poisson distribution model with $n = 100, p = 3, q = 1, k = 2$. The regression parameters and the mixed proportion were estimated using the proposed algorithm for $K = 1, 2, 3, 4, 5$. The BIC criterion was used to determine the optimal number of components and the corresponding BIC was depicted in Figure 4.2. The BIC was minimized at $K = 2$ which suggests 2 clusters. The proposed algorithm is able to detect the correct number of clusters. The resulting regression lines for each of the components separately were the black lines shown on Figure 4.3. To check if the two mixture components can be correctly identified, a cross-tabulation of true memberships and

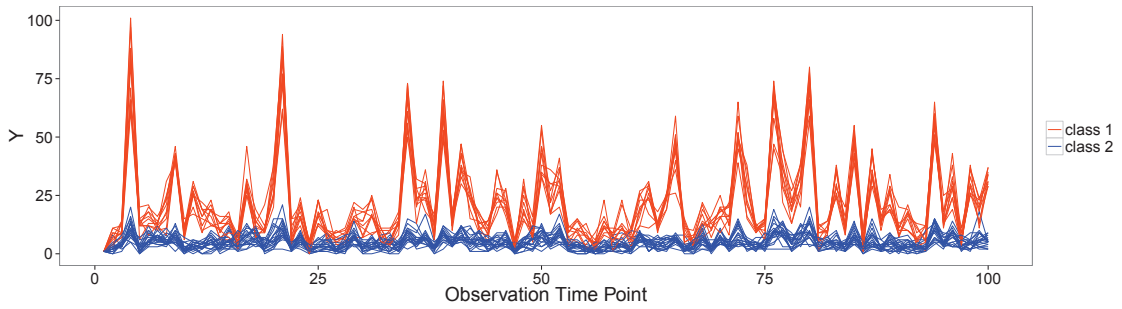

Figure 4.1 A two-component simulated data set of m=30 time series of size n=100.

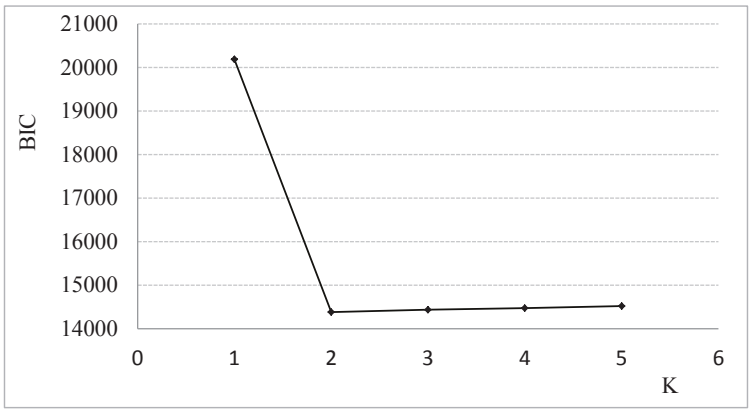

Figure 4.2 The BIC values for mixture models to the simulation data with components 1 to 5.

predicted memberships was shown in Table 4.1. It can be seen that the two mixture components were perfectly separated and the memberships were identified with great accuracy.

4.6 Clustering the mosquito surveillance data

In this section, we applied our algorithm to the mosquito surveillance data described in Chapter 2. The response variables $\boldsymbol{Y} = \{y_{ijt}\}$, $i = 1, \ldots, 29$, $j = 1, \ldots, 8$, where i indexes the culex mosquito counts from the 29 traps in Peel Region, j indexes the eight different years from 2004 to 2011, and t indexes weekly results. The same data smoothing technique used in Section 2.2.2 was applied to the mosquito data. The explanatory variables $\boldsymbol{X}_w$ were the weather data *ddm* and *ppm*. A mixture Markov regression model with conditional Poisson distribution was fitted to the mosquito time series. For the kth mixture component, the conditional mean of the mosquito counts in a trap conditional on the previous week counts is given by

$$\log(\boldsymbol{\mu}_{ijt}^{(k)}) = \boldsymbol{x}_{ijt}\boldsymbol{\beta}_k + \lambda_k y_{ij,t-1} \tag{4.97}$$

where $\boldsymbol{x}' = \{1, ddm, ppm, ddm \times ppm, ddm^2, ppm^2\}$. Note that because of the autoregressive term of $t-1$, the quasi-likelihood model will only sum the contributions from $t = 2$.

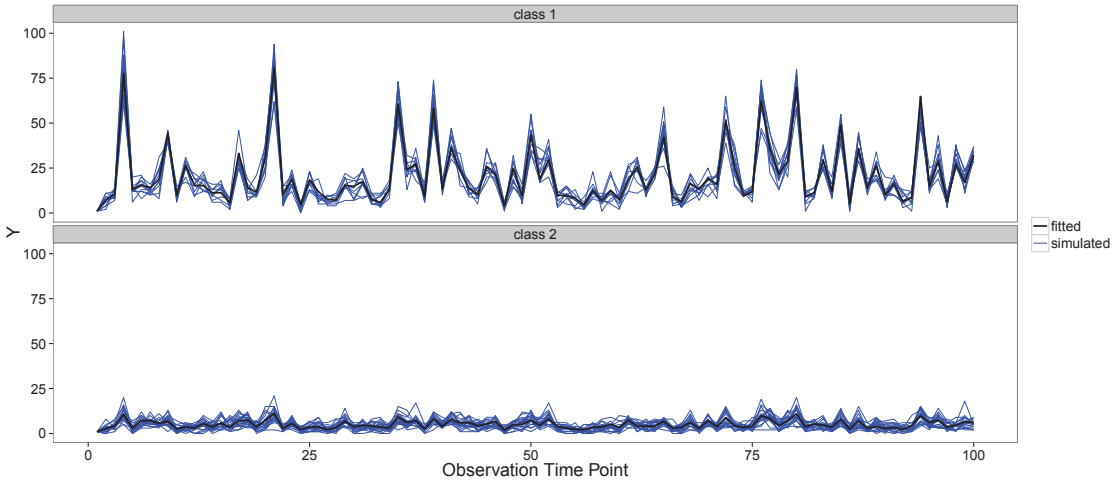

Figure 4.3 Fitted regression lines to the simulated data set with two-component.

Table 4.1 The cross-tabulation of true classes cluster memberships.

	class 1	class 2
cluster 1	10	0
cluster 2	0	20

We used R package FlexMix (Leisch 2004) to obtain the initial values for our algorithm. The optimal number of components was determined by the BIC criterion through fitting the model with different numbers of components ranging from 1 to 5. The corresponding BIC curve was depicted in Figure 4.4. The mixture model with 3 components achieved the smallest BIC and was selected as the optimal model.

We also examined the predicted probability for each trap to be assigned to the mixture components. For each trap, we assign the membership to the component with the maximum probability among the three components. For the mosquito surveillance data, the maximum predicted probabilities had a mean of 0.95 with a standard deviation of 0.11. This indicated that observations can be assigned to one of the components with high confidence and the components were well separated. The estimated regression parameters, the corresponding standard errors and estimated mixing proportions were listed in Table 4.2. The predicted mean values of mosquito counts for each component were depicted in Figure 4.5. It was evident that the different components responded differently to the weather changes.

We compared the results obtained from the proposed mixture Markov model and the hard clustering method which performs clustering based on K-means algorithm and build Markov regression model for each cluster. The RMSE of the two methods were calculated. The RMSE of the hard clustering method and the mix-

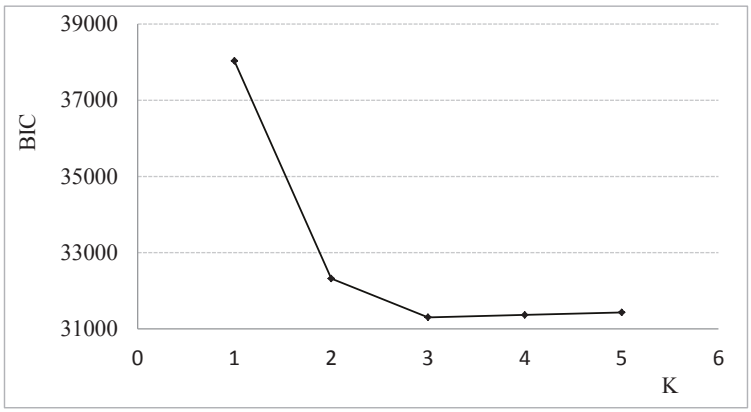

Figure 4.4 The BIC values for mixture models of the mosquito surveillance data with components 1 to 5.

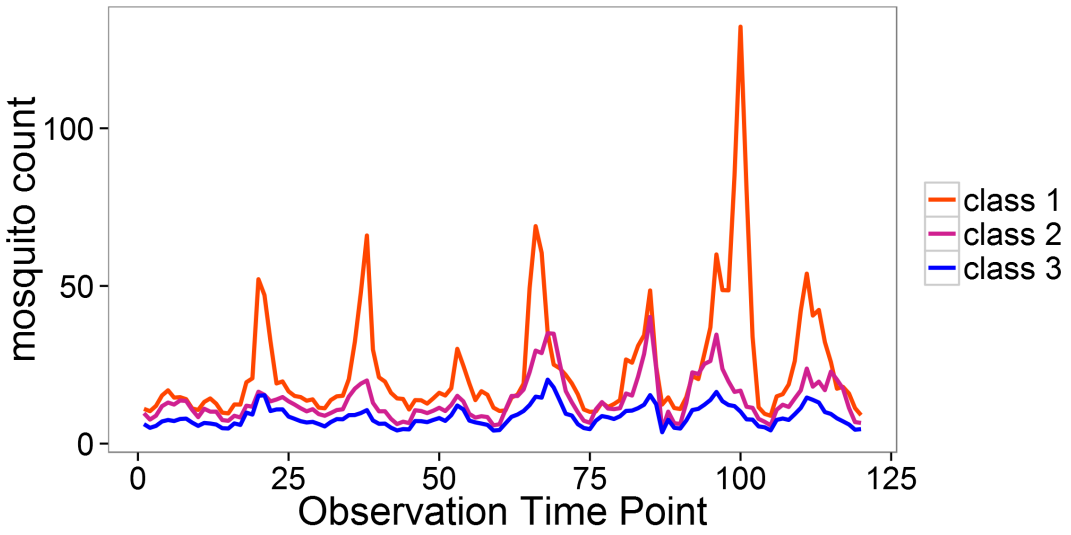

Figure 4.5 Mean mosquito traps patterns of the mixture Markov model fitted to the mosquito surveillance data in Peel Region.

Table 4.2 The estimated parameters and the corresponding standard error for the mosquito surveillance data with component equal to 3. For each cluster, the number in the bracket is the standard error of the corresponding parameter.

	β_0	β_1	β_2	β_3	β_4	β_5	θ	π
cluster 1	2.0293	-0.0168	-0.0579	0.0042	0.0089	0.0035	0.0174	0.042
	(0.2740)	(0.0088)	(0.0639)	(1.98e-05)	(2.03e-04)	(2.95e-04)	(1.42e-06)	
cluster 2	0.8815	0.0693	0.2665	0.0014	-0.0086	-0.0060	0.0198	0.280
	(0.0938)	(0.0028)	(0.0141)	(5.9542)	(4.40e-05)	(6.35e-05)	(2.86e-07)	
cluster 3	0.7464	0.0299	0.2846	0.0041	-0.0210	0.0077	0.0230	0.678
	(0.0572)	(0.0017)	(0.0101)	(3.74e-06)	(3.11e-05)	(4.64e-05)	(1.39e-07)	

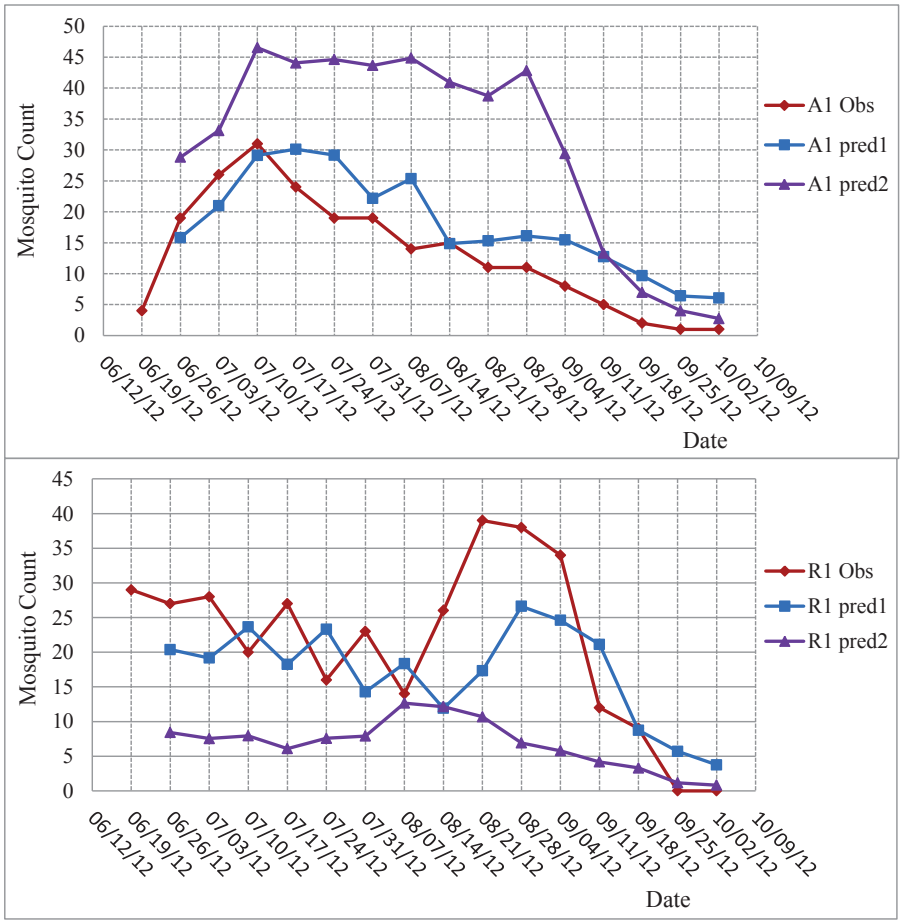

Figure 4.6 Predicting the mosquito abundance in 2012 by two clustering methods.

ture Markov model were 18.78 and 12.81 respectively. The mixture Markov model obtained smaller RMSE and therefore outperformed the hard clustering method in terms of model fitting. In this application, since the mosquito surveillance time series of each trap in Peel Region has same length, both hard clustering and the Markov mixture model can be applied. In the other health region of Ontario, the mosquito surveillance time series of each trap may vary in length. In this case, only the Markov mixture model works.

We also compared the two methods by predicting the mosquito abundance in Peel Region in 2012, where the two models were trained on data from 2004 to 2011. The RMSE of the hard clustering method and the mixture Markov model were 20.44 and 16.97 respectively. In Figure 4.6 displayed the prediction results of two traps by the two methods. It was evident that the prediction of the mixture Markov model was more accurate and the predicted curve by the mixture Markov model is much closer to the actual observations than the hard clustering method.

4.7 Conclusions

A general framework of finite mixture regression models with Markov process was proposed to analyze heterogeneous time series data. Quasi-likelihood method was proposed to estimate the parameters. A dedicated novel EM algorithm was devel-

oped to maximize the quasi-likelihood with mixture components. The asymptotic properties of the quasi-likelihood estimates were established. The proposed algorithm can be used to model and forecast heterogeneous time series data with exogenous variables. It can also be used as a quasi-likelihood based clustering algorithm for time series.

Table 4.3 The true parameters used in the simulation

	β	λ(L)	λ(H)	Pr
k=2 p=3 q=1	$\begin{pmatrix} 1.0 & 1.5 & 2.0 \\ 0.5 & 1.5 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.10 \\ 0.15 \end{pmatrix}$	$\begin{pmatrix} 0.60 \\ 0.65 \end{pmatrix}$	$\begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix}$
k=2 p=3 q=2	$\begin{pmatrix} 1.0 & 1.5 & 2.0 \\ 0.5 & 1.5 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.10 & 0.15 \\ 0.15 & 0.20 \end{pmatrix}$		$\begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix}$
k=2 p=5 q=1	$\begin{pmatrix} 1.0 & 1.5 & 2.0 & 0.5 & 0.2 \\ 0.5 & 1.5 & 0.5 & 1.0 & 0.3 \end{pmatrix}$	$\begin{pmatrix} 0.10 \\ 0.15 \end{pmatrix}$		$\begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix}$
k=3 p=3 q=1	$\begin{pmatrix} 1.0 & 1.5 & 2.0 \\ 0.5 & 1.5 & 0.5 \\ 0.5 & 1.0 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.10 \\ 0.15 \\ 0.20 \end{pmatrix}$		$\begin{pmatrix} 0.4 \\ 0.2 \\ 0.4 \end{pmatrix}$

Table 4.4 The estimated parameters and standard error for conditional Poisson distribution with $K = 2$. In each of the estimation, the first p parameters were β and the rest q ones were λ. SE-theo were the theoretical standard error and SE-simu were the standard error obtained from the simulation.

	k=1	k=2
n=100 p=3 q=1(L)		
est para	(1.001 1.500 1.999 0.101)	(0.502 1.498 0.499 0.150)
SE-theo	(0.042 0.029 0.030 0.010)	(0.038 0.037 0.035 0.011)
SE-simu	(0.030 0.030 0.031 0.037)	(0.032 0.037 0.036 0.015)
est prob	0.296	0.704
SE prob	0.083	0.083
n=50 p=3 q=1(L)		
est para	(1.001 1.500 2.000 0.102)	(0.498 1.502 0.501 0.150)
SE-theo	(0.060 0.042 0.044 0.013)	(0.055 0.053 0.051 0.015)
SE-simu	(0.044 0.046 0.030 0.047)	(0.049 0.033 0.053 0.022)
est prob	0.296	0.704
SE prob	0.081	0.081
n=50 p=3 q=1(H)		
est para	(1.000 1.500 2.000 0.596)	(0.496 1.500 0.500 0.648)
SE-theo	(0.137 0.039 0.042 0.016)	(0.110 0.054 0.051 0.018)
SE-simu	(0.057 0.038 0.040 0.041)	(0.070 0.051 0.050 0.021)
est prob	0.299	0.701
SE prob	0.082	0.082
n=50 p=3 q=2(L)		
est para	(0.988 1.498 1.958 0.099 0.152)	(0.512 1.500 0.541 0.150 0.197)
SE-theo	(0.082 0.043 0.045 0.014 0.014)	(0.070 0.055 0.053 0.016 0.017)
SE-simu	(0.094 0.045 0.246 0.060 0.046)	(0.098 0.055 0.278 0.025 0.025)
est prob	0.309	0.691
SE prob	0.107	0.107
n=50 p=5 q=1(L)		
est para	(0.999 1.500 2.001 0.501 0.199 0.099)	(0.499 1.499 0.502 1.000 0.300 0.149)
SE-theo	(0.060 0.036 0.038 0.035 0.034 0.011)	(0.058 0.039 0.037 0.038 0.037 0.014)
SE-simu	(0.046 0.036 0.039 0.037 0.035 0.043)	(0.045 0.037 0.037 0.038 0.036 0.026)
est prob	0.302	0.698
SE prob	0.089	0.089

Table 4.5 The estimated parameters and standard error for conditional Poisson distribution with $K = 3$. In each of the estimation, the first p parameters were β and the rest q ones were λ. SE-theo were the theoretical standard error and SE-simu were the standard error obtained from the simulation.

	k=1				k=2				k=3			
n=50 p=3 q=1(L)												
est para	1.001	1.498	2.000	0.098	0.503	1.496	0.497	0.148	0.499	1.000	0.502	0.200
SE-theo	0.051	0.036	0.037	0.011	0.108	0.104	0.099	0.032	0.075	0.061	0.062	0.020
SE-simu	0.034	0.035	0.036	0.039	0.098	0.108	0.110	0.046	0.058	0.058	0.062	0.033
est prob	0.401				0.201				0.398			
SE prob	0.088				0.072				0.085			

Table 4.6 The estimated parameters and standard error for conditional gamma distribution with $K = 2$. In each of the estimation, the first p parameters were β and the rest q ones were λ. SE-theo were the theoretical standard error and SE-simu were the standard error obtained from the simulation.

	k=1	k=2
n=50 p=3 q=1(L) $\psi = 1$		
est para	$\begin{pmatrix}0.996 & 1.496 & 1.998 & 0.097\end{pmatrix}$	$\begin{pmatrix}0.504 & 1.499 & 0.489 & 0.149\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.158 & 0.173 & 0.172 & 0.033\end{pmatrix}$	$\begin{pmatrix}0.090 & 0.110 & 0.109 & 0.024\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.147 & 0.186 & 0.186 & 0.044\end{pmatrix}$	$\begin{pmatrix}0.092 & 0.111 & 0.111 & 0.027\end{pmatrix}$
est prob	0.300	0.700
SE prob	0.086	0.086
n=100 p=3 q=1(L) $\psi = 1$		
est para	$\begin{pmatrix}0.997 & 1.502 & 2.002 & 0.100\end{pmatrix}$	$\begin{pmatrix}0.495 & 1.502 & 0.504 & 0.149\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.110 & 0.120 & 0.120 & 0.024\end{pmatrix}$	$\begin{pmatrix}0.063 & 0.076 & 0.076 & 0.017\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.099 & 0.126 & 0.127 & 0.030\end{pmatrix}$	$\begin{pmatrix}0.062 & 0.079 & 0.075 & 0.018\end{pmatrix}$
est prob	0.300	0.700
SE prob	0.084	0.084
n=100 p=3 q=1(L) $\psi = 3$		
est para	$\begin{pmatrix}0.995 & 1.502 & 1.995 & 0.100\end{pmatrix}$	$\begin{pmatrix}0.495 & 1.504 & 0.496 & 0.149\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.165 & 0.209 & 0.208 & 0.025\end{pmatrix}$	$\begin{pmatrix}0.104 & 0.133 & 0.132 & 0.019\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.177 & 0.211 & 0.227 & 0.028\end{pmatrix}$	$\begin{pmatrix}0.113 & 0.135 & 0.134 & 0.020\end{pmatrix}$
est prob	0.300	0.700
SE prob	0.086	0.086
n=100 p=3 q=1(H) $\psi = 1$		
est para	$\begin{pmatrix}0.998 & 1.500 & 2.003 & 0.599\end{pmatrix}$	$\begin{pmatrix}0.491 & 1.503 & 0.503 & 0.649\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.141 & 0.104 & 0.105 & 0.021\end{pmatrix}$	$\begin{pmatrix}0.078 & 0.065 & 0.065 & 0.015\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.142 & 0.106 & 0.107 & 0.024\end{pmatrix}$	$\begin{pmatrix}0.100 & 0.067 & 0.065 & 0.015\end{pmatrix}$
est prob	0.300	0.700
SE prob	0.083	0.083
n=100 p=3 q=2(L) $\psi = 1$		
est para	$\begin{pmatrix}0.997 & 0.499 & 1.999 & 0.099 & 0.149\end{pmatrix}$	$\begin{pmatrix}0.501 & 0.496 & 0.499 & 0.149 & 0.200\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.132 & 0.120 & 0.121 & 0.024 & 0.024\end{pmatrix}$	$\begin{pmatrix}0.068 & 0.075 & 0.075 & 0.017 & 0.017\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.109 & 0.127 & 0.124 & 0.033 & 0.031\end{pmatrix}$	$\begin{pmatrix}0.069 & 0.076 & 0.074 & 0.019 & 0.019\end{pmatrix}$
est prob	0.298	0.702
SE prob	0.084	0.084
n=100 p=5 q=1(L) $\psi = 1$		
est para	$\begin{pmatrix}1.002 & 1.496 & 2.001 & 0.499 & 0.194 & 0.099\end{pmatrix}$	$\begin{pmatrix}0.498 & 1.502 & 0.502 & 0.995 & 0.302 & 0.149\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.144 & 0.121 & 0.121 & 0.121 & 0.121 & 0.023\end{pmatrix}$	$\begin{pmatrix}0.088 & 0.078 & 0.077 & 0.078 & 0.077 & 0.016\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.135 & 0.122 & 0.122 & 0.122 & 0.130 & 0.031\end{pmatrix}$	$\begin{pmatrix}0.085 & 0.077 & 0.078 & 0.082 & 0.079 & 0.019\end{pmatrix}$
est prob	0.303	0.697
SE prob	0.083	0.083

Table 4.7 The estimated parameters and standard error for conditional gamma distribution model with $K = 3$. In each of the estimation, the first p parameters were β and the rest q ones were λ. SE-theo were the theoretical standard error and SE-simu were the standard error obtained from the simulation

	k=1				k=2				k=3			
n=50 p=3 q=1(L) $\psi = 1$												
est para	0.997	1.499	2.007	0.099	0.496	1.501	0.500	0.148	0.498	0.989	1.510	0.200
SE-theo	0.095	0.104	0.103	0.020	0.124	0.151	0.150	0.033	0.089	0.103	0.103	0.022
SE-simu	0.084	0.106	0.104	0.026	0.139	0.180	0.203	0.045	0.093	0.122	0.116	0.027
est prob	0.398				0.206				0.396			
SE prob	0.090				0.082				0.096			

Table 4.8 Binomial(a)The estimated parameters and standard error for conditional Binomial distribution model with $K = 2$. In each of the estimation, the first p parameters were β and the rest q ones were λ. SE-theo were the theoretical standard error and SE-simu were the standard error obtained from the simulation

	k=1	k=2
n=100 p=3 q=1(L)		
est para	$\begin{pmatrix}1.001 & 1.502 & 2.003 & 0.100\end{pmatrix}$	$\begin{pmatrix}0.499 & 1.501 & 0.501 & 0.150\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.079 & 0.041 & 0.053 & 0.011\end{pmatrix}$	$\begin{pmatrix}0.045 & 0.023 & 0.011 & 0.007\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.093 & 0.047 & 0.059 & 0.013\end{pmatrix}$	$\begin{pmatrix}0.046 & 0.024 & 0.012 & 0.007\end{pmatrix}$
est prob	0.301	0.699
SE prob	0.088	0.088
n=50 p=3 q=1(L)		
est para	$\begin{pmatrix}1.008 & 1.510 & 2.014 & 0.100\end{pmatrix}$	$\begin{pmatrix}0.504 & 1.503 & 0.500 & 0.150\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.115 & 0.059 & 0.075 & 0.016\end{pmatrix}$	$\begin{pmatrix}0.065 & 0.033 & 0.017 & 0.009\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.147 & 0.070 & 0.089 & 0.019\end{pmatrix}$	$\begin{pmatrix}0.069 & 0.036 & 0.018 & 0.010\end{pmatrix}$
est prob	0.298	0.702
SE prob	0.085	0.085
n=50 p=3 q=1(H)		
est para	$\begin{pmatrix}1.008 & 1.514 & 2.018 & 0.605\end{pmatrix}$	$\begin{pmatrix}0.510 & 1.504 & 0.502 & 0.651\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.162 & 0.068 & 0.086 & 0.031\end{pmatrix}$	$\begin{pmatrix}0.131 & 0.045 & 0.024 & 0.023\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.240 & 0.084 & 0.108 & 0.040\end{pmatrix}$	$\begin{pmatrix}0.170 & 0.053 & 0.028 & 0.029\end{pmatrix}$
est prob	0.298	0.702
SE prob	0.083	0.083
n=50 p=3 q=2(L)		
est para	$\begin{pmatrix}1.009 & 1.510 & 2.013 & 0.100 & 0.152\end{pmatrix}$	$\begin{pmatrix}0.505 & 1.503 & 0.501 & 0.151 & 0.200\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.163 & 0.062 & 0.080 & 0.018 & 0.018\end{pmatrix}$	$\begin{pmatrix}0.107 & 0.037 & 0.018 & 0.012 & 0.012\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.211 & 0.083 & 0.107 & 0.022 & 0.022\end{pmatrix}$	$\begin{pmatrix}0.123 & 0.051 & 0.022 & 0.013 & 0.013\end{pmatrix}$
est prob	0.298	0.702
SE prob	0.085	0.085
n=50 p=5 q=1(L)		
est para	$\begin{pmatrix}1.013 & 1.516 & 2.018 & 0.504 & 0.201 & 0.100\end{pmatrix}$	$\begin{pmatrix}0.496 & 1.505 & 0.501 & 1.003 & 0.301 & 0.151\end{pmatrix}$
SE-theo	$\begin{pmatrix}0.131 & 0.065 & 0.083 & 0.033 & 0.029 & 0.018\end{pmatrix}$	$\begin{pmatrix}0.072 & 0.037 & 0.019 & 0.027 & 0.016 & 0.010\end{pmatrix}$
SE-simu	$\begin{pmatrix}0.169 & 0.079 & 0.105 & 0.039 & 0.037 & 0.023\end{pmatrix}$	$\begin{pmatrix}0.076 & 0.039 & 0.020 & 0.028 & 0.018 & 0.011\end{pmatrix}$
est prob	0.298	0.702
SE prob	0.087	0.087

Table 4.9 The estimated parameters and standard error for conditional Binomial distribution model with $K = 3$. In each of the estimation, the first p parameters were β and the rest q ones were θ. SE-theo were the theoretical standard error and SE-simu were the standard error obtained from the simulation.

	k=1				k=2				k=3			
n=50 p=3 q=1(L)												
est para	1.004	1.508	2.010	0.101	0.491	1.507	0.502	0.152	0.501	1.004	1.506	0.201
SE-theo	0.104	0.053	0.068	0.014	0.127	0.065	0.032	0.018	0.094	0.034	0.048	0.014
SE-simu	0.116	0.064	0.083	0.016	0.140	0.073	0.035	0.021	0.101	0.037	0.051	0.014
est prob	0.401				0.201				0.398			
SE prob	0.095				0.074				0.093			

5 Conclusions and future work

The Fourth Assessment Report of the Intergovernmental Panel on Climate Changes (Parry et al. 2007) states that North America has experienced remarkable changes in climate patterns including warmer temperatures and increased rainfall, summertime droughts and abnormal weather events (e.g., tornadoes and hurricanes). Variations in weather can greatly affect the incidence and distribution of arthropod vectors such as mosquitoes. Researchers have long understood that weather conditions and environmental factors affect mosquito distribution, population sizes and their transmission of pathogens. Statistical models that provide the relationships between mosquito abundance and antecedent weather/environmental conditions are in great need for public health agencies to improve the efficiency of vector control. Thus in this dissertation, we tried to analyse the distribution properties of mosquito abundance in order to develop the forecasting models to predict the mosquito abundance under different weather conditions.

The first part of this thesis studied the distribution properties of mosquito abun-

dance using the data of *Culex pipens /restuans* species from the surveillance program in Peel Region, Ontario for the period of 2004 to 2012. K-means and agglomerative hierarchical clustering methods were combined to identify two clusters of mosquito traps that had similar data. Validation against landscape data suggested that the clusters represent different habitats, and that mosquitoes in different clusters had different capacity to change breeding rates in response to changing weather conditions. Accounting for the occurrence of these clusters, distribution analysis showed that *Culex* mosquito abundance in Peel Region followed a gamma distribution. However, mean summer temperature had a significant impact on the distribution properties: below a threshold mean summer temperature of 17.66C, the data were significantly different from a gamma distribution, and this signalled a year in which a double rather than single peak in mosquito abundance was observed. A predictive statistical model by clusters to forecast mosquito abundance using weather conditions was then developed. By using these methods of analysis to capture geographic variations, and threshold weather conditions in the response of *Culex* mosquito populations to changing weather, accurate weather-based forecasting was achieved.

In the second part of this dissertation, forecasting models were developed to predict the *Culex* mosquito abundance, the WNv risk and human incidence in GTA under weather changes. The weather conditions that affect the mosquito abundance

and WNv transmission were examined and the most significant temperature and precipitation were given in each case. Multiple models, such as Zero-Inflated Poisson (ZIP), gamma, Poisson, negative binomial and Zero-Inflated negative binomial distributions were proposed. Leave-one-out cross-validation, RMSE, and BIC, were used as the model selection criteria to choose the best fit models. The predictions were in a good agreement with the observations for the period from 2002 to 2012. The model selection was demonstrated to be an effective way to compare different models. These models chosen could be used by the public health authorities to forecast the WNv risk one week ahead.

In the last part, we introduced a new model-based clustering approach for time series. The proposed model consisted of finite mixture model govern by Markov process. Vector based clustering methods are insufficient to precess varying length data set. Quasi-likelihood approach was adopted to deal with the Markov chain in the data generating process. By using MLE, the parameters were estimated through EM algorithm and BIC was used to determined the optimal number of components. The proposed algorithm was tested on the simulated data set of conditional elliptical mixture models. The results demonstrated that algorithm can detect the correct number of clusters and the clusters were perfectly separated. The regression parameters estimated were very close to the true values. The algorithm was applied

to analyse the mosquito surveillance data in Peel Region. The traps in Peel Region were classified into 3 clusters. The estimated a-posteriori probabilities indicated that observations can be assigned to one of the 3 components with high confidence and the components were well separated. Comparison with the hard clustering method introduced in Chapter 2, showed that smaller error and more accurate prediction were achieved.

After the biggest outbreak of WNv in Ontario in 2002, the WNv activity has varied from year to year and started to decline from 2005 until 2010. The decreasing incidence rate of disease prior to 2010 reduced scientific interest in modeling the WNv transmission and some researchers (Adlouni et al. 2007) proposed a much scaled-down larvicide spraying and other mosquito control program due to the low probability of recurrence of the favourable climatic conditions like 2002. However, the emergence of WNv in Ontario in 2012 showed that WNv has already established itself in Ontario and could be reemergence more frequently. Although the models in this dissertation seem to be plausible, there are still a number of problems need to be discovered.

Firstly, the effects of these off-season conditions on mosquito population growth and WNv transmission have not been studied as extensively and the results have been mixed. Mogi (1996) have hypothesized that milder winters may lead to larger popu-

lations the following summer, while Wegbreit and Reisen (2000) demonstrated that higher levels of snow moisture and thus higher runoff from melting snow led to increases in the *Culex tarsalis* population. Some other off-season variables have shown to have correlations with mosquito abundance and WNv transmission (Adlouni et al. 2007, Reisen et al. 2008, Walsh et al. 2008). These results were not consistent and suggest that the effects of off-season conditions are complex. They probably also differ by species, particular since different species employ different strategies for surviving winter and rely on different cues to emerge. Including the off-season conditions into the forecasting models should improve the prediction accuracy.

Secondly, the existence of critical climate thresholds may affect the mosquito abundance and WNv transmission even without significant climate change (Patz et al. 2002). A deeper understanding of the relationship between climate and disease dynamics need to be done for anticipating the potential effects that a changing climate would have on the occurrence and distribution of the WNv.

Thirdly, in the application of finite mixture model, the same covariate may have a different impact on a different component. This creates a complex variable selection problem. The traditional selection methods such as Akaike Information Criterion (AIC) (Akaike 1973) and the BIC are computationally expensive for the mixture model. Tibshirani (1996) developed a new model selection technique called Least

Absolute Shrinkage and Selection Operator (LASSO). This algorithm deleted the non-significant covariates in the model by estimating their effects to be zero in the regression model and had favorable properties in model selection and interpretation. Khalili and Chen (2007) added a new class of weighted penalty function to the penalized likelihood and achieved LASSO-type simultaneous covariate selection and estimation. Liu et al. (2015) extended Bayesian approach by introducing component-adaptive weighted priors for regression coefficient, which can also simultaneously conducting covariate selection and determining the number of components. There is still a rather large search space in cooperating the above methods with our algorithm to obtain better model selection.

Finally, our proposed finite mixture model with Markov process has been applied to conditional Poisson, Gamma and binomial distributions successfully. As a future work, we wish to extend the proposed method to the setting that the conditional distribution may not come from an exponential family and only the first and second moments are specified. This imposes difficulty on imputing the missing memberships where there is neither likelihood nor conditional likelihood available. Finally, the model selection properties of BIC in the quasi-likelihood setting with mixture components need to be further investigated.

Bibliography

[1] Abeku, T. A., De Vlas, S. J., Borsboom, G., Teklehaimanot, A., Kebede, A., Olana, D., Van Oortmarssen, G. J., and Habbema, J. (2002). Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of ethiopia: a simple seasonal adjustment method performs best. *Tropical Medicine & International Health*, 7(10):851–857.

[2] Adlouni, S. E., Beaulieu, C., Ouarda, T. B., Gosselin, P. L., and Saint-Hilaire, A. (2007). Effects of climate on west nile virus transmission risk used for public health decision-making in quebec. *International journal of health geographics*, 6(1):40.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akadémiai Kiado.

[4] An, D. T. M. and Rocklöv, J. (2014). Epidemiology of dengue fever in hanoi from 2002 to 2010 and its meteorological determinants. *Global health action*, 7.

[5] Anderson, J. R. (1976). *A land use and land cover classification system for use with remote sensor data*, volume 964. US Government Printing Office.

[6] Anderson, R. L. (1954). The problem of autocorrelation in regression analysis. *Journal of the American Statistical Association*, 49(265):113–129.

[7] Artsob, H., Gubler, D., Enria, D., Morales, M., Pupo, M., Bunning, M., and Dudley, J. (2009). West nile virus in the new world: trends in the spread and proliferation of west nile virus in the western hemisphere. *Zoonoses and public health*, 56(6-7):357–369.

[8] Atashi, H., Sharbabak, M. M., and Shahrbabak, H. M. (2009). Environmental factors affecting the shape components of the lactation curves in holstein dairy cattle of iran. *Parity*, 2(354):4–66.

[9] Auld, H. and Service, C. A. E. (1990). *The climate of metropolitan Toronto*. Environment Canada, Atmospheric Environment Service.

[10] Azzalini, A. (1982). Approximate filtering of parameter driven processes. *Journal of Time Series Analysis*, 3(4):219–223.

[11] Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.

[12] Barker, C., Reisen, W., Eldridge, B., Park, B., and Johnson, W. (2009). Culex tarsalis abundance as a predictor of western equine encephalomyelitis virus transmission. *Proc. Mosq. Vector Control Assoc. Calif*, 77:65–68.

[13] Basford, K. E. and McLachlan, G. J. (1985). The mixture method of clustering applied to three-way data. *Journal of Classification*, 2(1):109–125.

[14] Bernkopf, H., Levine, S., and Nerson, R. (1953). Isolation of west nile virus in israel. *Journal of infectious diseases*, 93(3):207–218.

[15] Blender, R., Fraedrich, K., and Lunkeit, F. (1997). Identification of cyclone-track regimes in the north atlantic. *Quarterly Journal of the Royal Meteorological Society*, 123(539):727–741.

[16] Bolling, B. G., Kennedy, J. H., and Zimmerman, E. G. (2005). Seasonal dynamics of four potential west nile vector species in north-central texas. *Journal of Vector Ecology*, 30(2):186.

[17] Brown, H., Diuk-Wasser, M., Andreadis, T., and Fish, D. (2008). Remotely-sensed vegetation indices identify mosquito clusters of west nile virus vectors in an urban landscape in the northeastern united states. *Vector-Borne and Zoonotic Diseases*, 8(2):197–206.

[18] Brownstein, J. S., Rosen, H., Purdy, D., Miller, J. R., Merlino, M., Mostashari, F., and Fish, D. (2002). Spatial analysis of west nile virus: rapid risk assessment of an introduced vector-borne zoonosis. *Vector Borne and Zoonotic Diseases*, 2(3):157–164.

[19] Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2000). Visualization of navigation patterns on a web site using model-based clustering. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–284. ACM.

[20] Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery*, 7(4):399–424.

[21] Chamroukhi, F., Samé, A., Aknin, P., and Govaert, G. (2011). Model-based clustering with hidden markov model regression for time series with regime changes. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2814–2821. IEEE.

[22] Chuang, T.-W., Hildreth, M. B., Vanroekel, D. L., and Wimberly, M. C. (2011). Weather and land cover influences on mosquito populations in sioux falls, south dakota. *Journal of medical entomology*, 48(3):669–679.

[23] Cooke, W. H., Grala, K., and Wallis, R. C. (2006). Avian gis models signal human risk for west nile virus in mississippi. *International Journal of Health Geographics*, 5(1):36.

[24] Costa, A. C. C., Codeço, C. T., Honório, N. A., Pereira, G. R., Pinheiro, C. F. N., and Nobre, A. A. (2015). Surveillance of dengue vectors using spatio-temporal bayesian modeling. *BMC medical informatics and decision making*, 15(1):93.

[25] Costantino, R. and Desharnais, R. (1981). Gamma distributions of adult numbers for tribolium populations in the region of their steady states. *The Journal of Animal Ecology*, pages 667–681.

[26] Cox, D. R. (1970). *Analysis of binary data*. Chapman and Hall, London.

[27] Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K., and Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 93–115.

[28] DeGaetano, A. T. (2005). Meteorological effects on adult mosquito (culex) populations in metropolitan new jersey. *International Journal of Biometeorology*, 49(5):345–353.

[29] DeGroote, J., Sugumaran, R., Brend, S., Tucker, B., and Bartholomay, L. (2008). Landscape, demographic, entomological, and climatic associations with human disease incidence of west nile virus in the state of iowa, usa. *International Journal of Health Geographics*, 7(1):19.

[30] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

[31] DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, 5(2):249–282.

[32] Descloux, E., Mangeas, M., Menkes, C. E., Lengaigne, M., Leroy, A., Tehei, T., Guillaumot, L., Teurlai, M., Gourinat, A.-C., Benzler, J., et al. (2012). Climate-based models for understanding and forecasting dengue epidemics. *PLoS neglected tropical diseases*, 6(2):e1470.

[33] Diuk-Wasser, M. A., Brown, H. E., Andreadis, T. G., and Fish, D. (2006). Modeling the spatial distribution of mosquito vectors for west nile virus in connecticut, usa. *Vector-Borne & Zoonotic Diseases*, 6(3):283–295.

[34] Eickeler, S., Wallhoff, F., Lurgel, U., and Rigoll, G. (2001). Content based indexing of images and video using face detection and recognition methods. In *Acoustics,*

Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on, volume 3, pages 1505–1508. IEEE.

[35] Epstein, P. R. (2001). West nile virus and the climate. *Journal of Urban Health*, 78(2):367–371.

[36] Fahrmeir, L. (1989). Extended kalman filtering for nonnormal longitudinal data. In *Statistical Modelling*, pages 151–156. Springer.

[37] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

[38] Fuller, W. A. (2009). *Introduction to statistical time series*, volume 428. John Wiley & Sons.

[39] Gaffney, S. and Smyth, P. (1999). Trajectory clustering with mixtures of regression models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 63–72. ACM.

[40] Gómez, A., Kilpatrick, A. M., Kramer, L. D., Dupuis, A. P., et al. (2008). Land use and west nile virus seroprevalence in wild mammals. *Emerging infectious diseases*, 14(6):962.

[41] Gough, W. (2000). A climate perspective on the blizzard of99. *CMOS Bull*, 28:17.

[42] Grün, B. and Leisch, F. (2008). Finite mixtures of generalized linear regression models. In *Recent advances in linear models and related areas*, pages 205–230. Springer.

[43] Gu, W., Lampman, R., and Novak, R. J. (2003). Problems in estimating mosquito infection rates using minimum infection rate. *Journal of medical entomology*, 40(5):595–596.

[44] Harding, J. (1949). The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biological Association of the United Kingdom*, 28(01):141–153.

[45] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108.

[46] Hayes, E. B., Komar, N., Nasci, R. S., Montgomery, S. P., O'Leary, D. R., and Campbell, G. L. (2005). Epidemiology and transmission dynamics of west nile virus disease. *Emerging infectious diseases*, 11(8).

[47] Hirano, S. S., Nordheim, E. V., Arny, D. C., and Upper, C. D. (1982). Lognormal distribution of epiphytic bacterial populations on leaf surfaces. *Applied and Environmental Microbiology*, 44(3):695–700.

[48] Hu, W., Nicholls, N., Lindsay, M., Dale, P., McMICHAEL, A. J., Mackenzie, J. S., and Tong, S. (2004). Development of a predictive model for ross river virus disease in brisbane, australia. *The American journal of tropical medicine and hygiene*, 71(2):129–137.

[49] Hu, W., Tong, S., Mengersen, K., and Oldenburg, B. (2006). Rainfall, mosquito density and the transmission of ross river virus: A time-series forecasting model. *Ecological modelling*, 196(3):505–514.

[50] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233.

[51] Kalbfleisch, J. and Lawless, J. F. (1985). The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, 80(392):863–871.

[52] Kaufmann, H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory. *The Annals of Statistics*, pages 79–98.

[53] Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479).

[54] Kilpatrick, A. M., Kramer, L. D., Campbell, S. R., Alleyne, E. O., Dobson, A. P.,

Daszak, P., et al. (2005). West nile virus risk assessment and the bridge vector paradigm. *Emerg Infect Dis*, 11(3):425–429.

[55] Koenraadt, C. J. M. and Harrington, L. (2008). Flushing effect of rain on container-inhabiting mosquitoes aedes aegypti and culex pipiens (diptera: Culicidae). *Journal of medical entomology*, 45(1):28–35.

[56] Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

[57] Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies ii. clustering systems. *The computer journal*, 10(3):271–277.

[58] Lanciotti, R., Roehrig, J., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K., Crabtree, M., Scherret, J., et al. (1999). Origin of the west nile virus responsible for an outbreak of encephalitis in the northeastern united states. *Science*, 286(5448):2333–2337.

[59] Landesman, W., Allan, B., Langerhans, R., Knight, T., and Chase, J. (2007). Internanual associations between precipitation and human incidence of West Nile virus in the United States. *Vector- Borne and Zoonotics Diseases*, 7:337–343.

[60] Lawrance, A. and Lewis, P. (1985). Modelling and residual analysis of nonlinear

autoregressive time series in exponential variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 165–202.

[61] Lee, A. H., Wang, K., and Yau, K. K. (2001). Analysis of zero-inflated poisson data incorporating extent of exposure. *Biometrical Journal*, 43(8):963–975.

[62] Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent glass regression in r.

[63] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, pages 13–22.

[64] Liao, T. W. (2005). Clustering of time series dataa survey. *Pattern recognition*, 38(11):1857–1874.

[65] Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.

[66] Lindquist, A., Ikeshoji, T., Grab, B., De Meillon, B., and Khan, Z. (1967). Dispersion studies of culex pipiens fatigans tagged with 32 p in the kemmendine area of rangoon, burma. *Bulletin of the World Health Organization*, 36(1):21.

[67] Liu, W., Zhang, B., Zhang, Z., Tao, J., and Branscum, A. J. (2015). Model selection

in finite mixture of regression models: a bayesian approach with innovative weighted g priors and reversible jump markov chain monte carlo implementation. *Journal of Statistical Computation and Simulation*, 85(12):2456–2478.

[68] Madder, D., Surgeoner, G., and Helson, B. (1983). Number of generations, egg production, and developmental time of culex pipiens and culex restuans (diptera: Culicidae) in southern ontario. *Journal of medical entomology*, 20(3):275–287.

[69] Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (2008). *Forecasting methods and applications*. John Wiley & Sons.

[70] Marcantonio, M., Rizzoli, A., Metz, M., Rosà, R., Marini, G., Chadwick, E., and Neteler, M. (2015). Identifying the environmental conditions favouring west nile virus outbreaks in europe. *PloS one*, 10(3):e0121158.

[71] Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8(11):1235–1246.

[72] CDC (2007). Mosquito-borne diseases. Available at: `http://www.cdc.gov/ncidod/diseases/list_mosquitoborne.htm`.

[73] CDC (2013). Morbidity and mortality weekly report (mmwr). *Centers for Diease Control and Prevention*. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6225a1.htm.

[74] CDC (2015). West nile virus (wnv) fact sheet. *Centers for Diease Control and Prevention*. Available at: http://www.cdc.gov/westnile/resources/pdfs/wnvFactsheet_508.pdf.

[75] EC (2012). Ontario weather review 2012. *Environmental Canada*. Available at: http://www.ec.gc.ca/default.asp?lang=En&n=DE75E50D-1.

[76] Health Canada (2008). Canada's health concerns from climate change and variability. *Ottawa, Ont: Health Canada; 2007*. Available at: www.hc-sc.gc.ca/ewh-semt/altformats/hecssesc/pdf/climat/healthtable-tableausante.pdf (accessed 2008 Jan).

[77] MOHLTC (2008). West nile virus preparedness and prevention plan 2008. *Ontario Ministry of Health and Long-Term Care*. Available at: http://www.health.gov.on.ca/en/common/ministry/publications/reports/wnv_plan_2008/wnv_plan_full.pdf.

[78] MOHLTC (2013). Vector-borne diseases 2012 summary report. *Ontario Agency for Health Protection and Promotion (Public Health Ontario). Toronto, ON: Queen's*

Printer for Ontario. Available at: http://www.publichealthontario.ca/en/eRepository/Vector_Borne_Diseases_Summary_Report_2012.pdf.

[79] NLCD (1992). Nlcd 92 land cover class definitions. *U.S. Environmental Protection Agency*. Available at: http://www.epa.gov/mrlc/nlcd.html.

[80] NRC (2007). Health effects of climate change and climate variability. *Natural Resources Canada*. Available at: http://www.nrcan.gc.ca/environment/resources/publications/impacts-adaptation/reports/assessments/2004/ch9/10225.

[81] PHAC (2015). West nile virus monitor. *Public Health Agency of Canada*. Public Health Agency of Canada, Available at: http://healthycanadians.gc.ca/diseases-conditions-maladies-affections/disease-maladie/west-nile-nil-occidental/index-eng.php.

[82] PHO (2012). Vector borne diseases 2011 summary report. *Public Health Ontario*. Available at: http://www.publichealthontario.ca/en/DataAndAnalytics/Documents/PHO_Vector_Borne_Disease_Report_2011_June_26_2012_Final.pdf.

[83] PPH (2002). West nile virus in region of peel 2002. *Peel Public Health*. Available at: http://www.peelregion.ca/health/vbd/resources/reports.htm.

[84] PPH (2008). 2008 west nile virus in the region of peel - technical report. *Peel Public Health*. Available at: `http://www.peelregion.ca/health/vbd/pdfs/2008-wnv-tech-report.pdf`.

[85] Statistic Canada (2012). 2011 census: Population and dwelling counts. Available at: `http://www1.toronto.ca/city_of_toronto/social_development_finance_administration/files/pdf/2011-census-backgrounder.pdf`.

[86] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. London England Chapman and Hall 1983.

[87] McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

[88] Milligan, G. W. and Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied psychological measurement*, 11(4):329–354.

[89] Mogi, M. (1996). Overwintering strategies of mosquitoes (diptera: Culicidae) on warmer islands may predict impact of global warming on kyushu, japan. *Journal of medical entomology*, 33(3):438–444.

[90] Moore, C. G., McLean, R., Mitchell, C. J., Nasci, R., Tsai, T., Calisher, C., Marfin, A., Moore, P., and Gubler, D. (1993). *Guidelines for arbovirus surveillance programs in the United States*, volume 500. Department of Health and Human Services. Division of Vector-Borne Infectious Diseases.

[91] Murray, K. O., Ruktanonchai, D., Hesalroad, D., Fonken, E., and Nolan, M. S. (2013). West nile virus, texas, usa, 2012. *Emerging infectious diseases*, 19(11):1836.

[92] Nelder, J. A. and Baker, R. (1972). Generalized linear models. *Encyclopedia of Statistical Sciences*.

[93] Nelms, B. M., Macedo, P. A., Kothera, L., Savage, H. M., and Reisen, W. K. (2013). Overwintering biology of culex (diptera: Culicidae) mosquitoes in the sacramento valley of california. *Journal of medical entomology*, 50(4):773–790.

[94] Parry, M., Canziani, O., Palutikof, J., van der Linden, P., and Hanson, C., editors (2007). *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

[95] Patz, J. A., Epstein, P. R., Burke, T. A., and Balbus, J. M. (1996). Global climate change and emerging infectious diseases. *Jama*, 275(3):217–223.

[96] Patz, J. A., Hulme, M., Rosenzweig, C., Mitchell, T. D., Goldberg, R. A., Githeko, A. K., Lele, S., McMichael, A. J., and Le Sueur, D. (2002). Climate change (communication arising): Regional warming and malaria resurgence. *Nature*, 420(6916):627–628.

[97] Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, pages 71–110.

[98] Pecoraro, H. L., Day, H. L., Reineke, R., Stevens, N., Withey, J. C., Marzluff, J. M., and Meschke, J. S. (2007). Climatic and landscape correlates for potential west nile virus mosquito vectors in the seattle region. *Journal of Vector Ecology*, 32(1):22–28.

[99] Poulsen, C. S. (1990). Mixed markov and latent markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing*, 7(1):5–19.

[100] Pradier, S., Leblond, A., and Durand, B. (2008). Land cover, landscape structure, and west nile virus circulation in southern france. *Vector-Borne and Zoonotic Diseases*, 8(2):253–264.

[101] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[102] Rabiner, L. R., Lee, C., Juang, B., and Wilpon, J. (1989). Hmm clustering for connected word recognition. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 405–408. IEEE.

[103] Raddatz, R. (1986). A biometeorological model of an encephalitis vector. *Boundary-Layer Meteorology*, 34(1-2):185–199.

[104] Reeves, W. C., Hardy, J. L., Reisen, W. K., and Milby, M. M. (1994). Potential effect of global warming on mosquito-borne arboviruses. *Journal of medical entomology*, 31(3):323–332.

[105] Reisen, W. and Brault, A. C. (2007). West nile virus in north america: perspectives on epidemiology and intervention. *Pest management science*, 63(7):641–646.

[106] Reisen, W. K. (1995). Effect of temperature on culex tarsalis (diptera: Culicidae) from the coachella and san joaquin valleys of california. *Journal of medical entomology*, 32(5):636–645.

[107] Reisen, W. K., Cayan, D., Tyree, M., Barker, C. M., Eldridge, B., and Dettinger, M. (2008). Impact of climate variation on mosquito abundance in california. *Journal of vector ecology*, 33(1):89–98.

[108] Reisen, W. K., Fang, Y., and Martinez, V. M. (2006). Effects of temperature on the transmission of west nile virus by culex tarsalis (diptera: Culicidae). *Journal of medical entomology*, 43(2):309–317.

[109] Rosà, R., Marini, G., Bolzoni, L., Neteler, M., Metz, M., Delucchi, L., Chadwick, E. A., Balbo, L., Mosca, A., Giacobini, M., et al. (2014). Early warning of west nile virus mosquito vector: climate and land use models successfully explain phenol-

ogy and abundance of culex pipiens mosquitoes in north-western italy. *Parasites & vectors*, 7(1):1.

[110] Ruiz, M. O., Chaves, L. F., Hamer, G. L., Sun, T., Brown, W. M., Walker, E. D., Haramis, L., Goldberg, T. L., and Kitron, U. D. (2010). Local impact of temperature and precipitation on west nile virus infection in culex species mosquitoes in northeast illinois, usa. *Parasit Vectors*, 3(1):19.

[111] Schliep, A., Schönhuth, A., and Steinhoff, C. (2003). Using hidden markov models to analyze gene expression time course data. *Bioinformatics*, 19(suppl 1):i255–i263.

[112] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[113] Sclove, S. (2001). Note on clustering analysis. Available at: `https://www.uic.edu/classes/idsc/ids472/clustering.htm`.

[114] Service, M. (1993). *Mosquito Ecology: Field sampling methods*. London, UK: Chapman and Hall.

[115] Shelton, R. M. (1973). The effect of temperatures on development of eight mosquito species. *Mosq. News*, 33(1):1–12.

[116] Shone, S. M., Curriero, F. C., Lesser, C. R., and Glass, G. E. (2006). Characterizing

population dynamics of aedes sollicitans (diptera: Culicidae) using meteorological data. *Journal of medical entomology*, 43(2):393–402.

[117] Simoes, T. C., Codeço, C. T., Nobre, A. A., and Eiras, A. E. (2013). Modeling the non-stationary climate dependent temporal dynamics of aedes aegypti. *PLoS One*, 8(8):e64773.

[118] Singh, A. K., Singh, A., and Engelhardt, M. (1997). The lognormal distribution in environmental applications. *Washington, DC: US Environmental Protection Agency, Office of Solid Waste and Emergency Response*.

[119] Smithburn, K., Hughes, T., Burke, A., Paul, J., et al. (1940). A neurotropic virus isolated from the blood of a native of uganda. *American Journal of Tropical Medicine*, 20:471–2.

[120] Steinbach, M., Karypis, G., Kumar, V., et al. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston.

[121] Sutherst, R. W. (2004). Global change and human vulnerability to vector-borne diseases. *Clinical microbiology reviews*, 17(1):136–173.

[122] Tibbetts, J. (2007). Driven to extremes health effects of climate change. *Environmental health perspectives*, 115(4):A196.

[123] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[124] Trawinski, P. and Mackay, D. (2008). Meteorologically conditioned time-series predictions of west nile virus vector mosquitoes. *Vector-Borne and Zoonotic Diseases*, 8(4):505–522.

[125] Turell, M. J., Dohm, D. J., Sardelis, M. R., O'guinn, M. L., Andreadis, T. G., and Blow, J. A. (2005). An update on the potential of north american mosquitoes (diptera: Culicidae) to transmit west nile virus. *Journal of medical entomology*, 42(1):57–62.

[126] Walsh, A. S., Glass, G. E., Lesser, C. R., and Curriero, F. C. (2008). Predicting seasonal abundance of mosquitoes based on off-season meteorological conditions. *Environmental and Ecological Statistics*, 15(3):279–291.

[127] Walton, M. (2010). Profile of agricultural attributes in the gta, phase 2. *Planscape Inc.* Available at: `http://www.planscape.ca/planscapePDFs/42-plan3.pdf`.

[128] Wang, J., Ogden, N. H., and Zhu, H. (2011). The impact of weather conditions on culex pipiens and culex restuans (diptera: Culicidae) abundance: a case study in peel region. *Journal of medical entomology*, 48(2):468–475.

[129] Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika*, 61(3):439–447.

[130] Wedel, M. and Kamakura, W. A. (2012). *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media.

[131] Wegbreit, J. and Reisen, W. K. (2000). Relationships among weather, mosquito abundance, and encephalitis virus activity in california: Kern county 1990-98. *Journal of the American Mosquito Control Association*, 16(1):22–27.

[132] West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83.

[133] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

[134] Wilding, G. E. and Mudholkar, G. S. (2008). A gamma goodness-of-fit test based on characteristic independence of the mean and coefficient of variation. *Journal of Statistical Planning and Inference*, 138(12):3813–3821.

[135] Wimberly, M. C., Hildreth, M. B., Boyte, S. P., Lindquist, E., and Kightlinger, L.

(2008). Ecological niche of the 2003 west nile virus epidemic in the northern great plains of the united states. *PLoS one*, 3(12):e3744.

[136] Xiong, Y. and Yeung, D.-Y. (2004). Time series clustering with arma mixtures. *Pattern Recognition*, 37(8):1675–1689.

[137] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987.

[138] Yoo, E.-H., Chen, D., Diao, C., and Russell, C. (2016). The effects of weather and environmental factors on west nile virus mosquito abundance in greater toronto area. *Earth Interactions*, 20(3):1–22.

[139] Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, pages 1019–1031.